VOLUME 102 PART B NUMBER 6 NOVEMBER 1955

The Proceedings

OF

THE INSTITUTION OF ELECTRICAL ENGINEERS

FOUNDED 1871: INCORPORATED BY ROYAL CHARTER 1921

PART B

RADIO AND ELECTRONIC ENGINEERING

(INCLUDING COMMUNICATION ENGINEERING)

The Institution of Electrical Engineers

FOUNDED 1871
INCORPORATED BY ROYAL CHARTER 1921

PATRON: HER MAJESTY THE QUEEN

COUNCIL 1955-56

President
SIR GEORGE H. NELSON, Bart.

Past-Presidents
SIR JAMES SWINBURNE, Bart., F.R.S.
W. H. ECCLES, D.Sc., F.R.S.
THE RT. HON. THE EARL OF MOUNT EDGCUMBE, T.D.
J. M. DONALDSON, M.C.
PROFESSOR E. W. MARCHANT, D.Sc.
P. V. HUNTER, C.B.E.
H. T. YOUNG.
SIR GEORGE LEE, O.B.E., M.C.
SIR ARTHUR P. M. FLEMING, C.B.E., D.Eng., LL.D.
J. R. BEARD, C.B.E., M.Sc.
SIR NOEL ASHBRIDGE, B.Sc.(Eng.).
COLONEL SIR A. STANLEY ANGWIN, K.B.E., D.S.O., M.C., T.D., D.Sc. (Eng.).
SIR HARRY RAILING, D.Eng.
P. DUNSHEATH, C.B.E., M.A., D.Sc. (Eng.).
SIR VINCENT Z. DE FERRANTI, M.C.
T. G. N. HALDANE, M.A.
PROFESSOR E. B. MOULLIN, M.A., Sc.D.
SIR ARCHIBALD J. GILL, B.Sc.(Eng.).
SIR JOHN HACKING.
COLONEL B. H. LEESON, C.B.E., T.D.
SIR HAROLD BISHOP, C.B.E.,B.Sc.(Eng.).
J. ECCLES, C.B.E., B.Sc.

Vice-Presidents
T. E. GOLDUP, C.B.E.
S. E. GOODALL, M.Sc.(Eng.).
WILLIS JACKSON, D.Sc., D.Phil., F.R.S.
SIR HAMISH D. MACLAREN, K.B.E., C.B., D.F.C., LL.D., B.Sc.
SIR W. GORDON RADLEY, C.B.E., Ph.D.(Eng.).

Honorary Treasurer
THE RT. HON. THE VISCOUNT FALMOUTH.

Ordinary Members of Council
PROFESSOR H. E. M. BARLOW, Ph.D., B.Sc.(Eng.).
J. BENNETT.
C. M. COCK.
A. R. COOPER, M.Eng.
A. T. CRAWFORD, B.Sc.
B. DONKIN, B.A.
PROFESSOR J. GREIG, M.Sc., Ph.D.
F. J. LANE, O.B.E., M.Sc.
G. S. C. LUCAS, O.B.E.
D. MCDONALD, B.Sc.
C. T. MELLING, C.B.E., M.Sc.Tech.
H. H. MULLENS, B.Sc.
W. F. PARKER.
R. L. SMITH-ROSE, C.B.E., D.Sc., Ph.D.
G. L. WATES, J.P.
G. O. WATSON.
D. B. WELBOURN, M.A.
J. H. WESTCOTT, B.Sc.(Eng.), Ph.D.
E. L. E. WHEATCROFT, M.A.
R. T. B. WYNN, C.B.E., M.A.

Chairmen and Past-Chairmen of Sections
Measurement and Control:
W. BAMFORD, B.Sc.
*M. WHITEHEAD.

Radio and Telecommunication:
H. STANESBY.
*C. W. OATLEY, M.A., M.Sc.

Supply:
L. DRUCQUER.
*J. D. PEATTIE, B.Sc.

Utilization:
D. B. HOGG, M.B.E., T.D.
*J. I. BERNARD, B.Sc.Tech.

Chairmen and Past-Chairmen of Local Centres
East Midland Centre:
F. R. C. ROBERTS.
*J. M. MITCHELL, B.Sc., Ph.D.

Mersey and North Wales Centre:
PROFESSOR J. M. MEEK, D.Eng.
*P. R. DUNN, B.Sc.

North Midland Centre:
F. BARRELL.
*G. CATON.

North-Eastern Centre:
A. H. KENYON.
*G. W. B. MITCHELL, B.A.

Southern Centre:
L. H. FULLER, B.Sc.(Eng.).
*E. A. LOGAN, M.Sc.

North-Western Centre:
G. V. SADLER.
*PROFESSOR E. BRADSHAW, M.B.E., M.Sc.Tech., Ph.D.

Northern Ireland Centre:
MAJOR E. N. CUNLIFFE, B.Sc.Tech.
*MAJOR P. L. BARKER, B.Sc.

Western Centre:
T. G. DASH, J.P.
*A. N. IRENS.

Scottish Centre:
*E. WILKINSON, Ph.D., B.Eng.
*J. S. HASTIE, B.Sc.(Eng.).

South Midland Centre:
H. S. DAVIDSON, T.D.
*A. R. BLANDFORD.

* Past-Chairman.

RADIO AND TELECOMMUNICATION SECTION COMMITTEE 1955-56

Chairman
H. STANESBY.

Vice-Chairmen
R. C. G. WILLIAMS, Ph.D., B.Sc.(Eng.).
J. S. MCPETRIE, Ph.D., D.Sc.

Past-Chairmen
C. W. OATLEY, M.A., M.Sc.
J. A. SMALE, C.B.E., A.F.C., B.Sc.

Ordinary Members of Committee
PROF. H. E. M. BARLOW, Ph.D., B.Sc.(Eng.).
F. S. BARTON, C.B.E., M.A., B.Sc.
A. J. BIGGS, Ph.D., B.Sc.
G. G. MACFARLANE, Dr.Ing., B.Sc.
B. N. MACLARTY, O.B.E.
H. PAGE, M.Sc.
N. C. ROBERTSON, C.M.G., M.B.E.
W. ROSS, M.A.
L. RUSHFORTH, M.B.E., B.Sc.
T. B. D. TERRONI, B.Sc.
A. M. THORNTON, B.Sc.
F. WILLIAMS, B.Sc.

And

The President (*ex officio*).
The Chairman of the Papers Committee.
PROF. H. E. M. BARLOW, Ph.D., B.Sc.(Eng.) (representing the Council).
E. H. COOKE-YARBOROUGH (Co-opted member).
BRIG. E. J. H. MOPPETT (representing the Cambridge Radio and Telecommunication Group).
J. MOIR (representing the South Midland Radio and Telecommunication Group).
D. H. THOMAS, M.Sc.Tech., B.Sc.(Eng.) (representing the North-Eastern Radio and Measurements Group).

The following nominees of Government Departments:
Admiralty: CAPTAIN G. C. F. WHITAKER, R.N.
Air Ministry: AIR COMMODORE G. H. RANDLE, R.A.F., B.A.
Department of Scientific and Industrial Research: B. G. PRESSEY, M.Sc.(Eng.), Ph.D.
Ministry of Supply: BRIG. J. D. HAIGH, O.B.E., M.A.
Post Office: CAPTAIN C. F. BOOTH, O.B.E.
War Office: COL. E. I. E. MOZLEY, M.A.

Secretary
W. K. BRASHER, C.B.E., M.A., M.I.E.E.

Deputy Secretary
F. JERVIS SMITH, M.I.E.E.

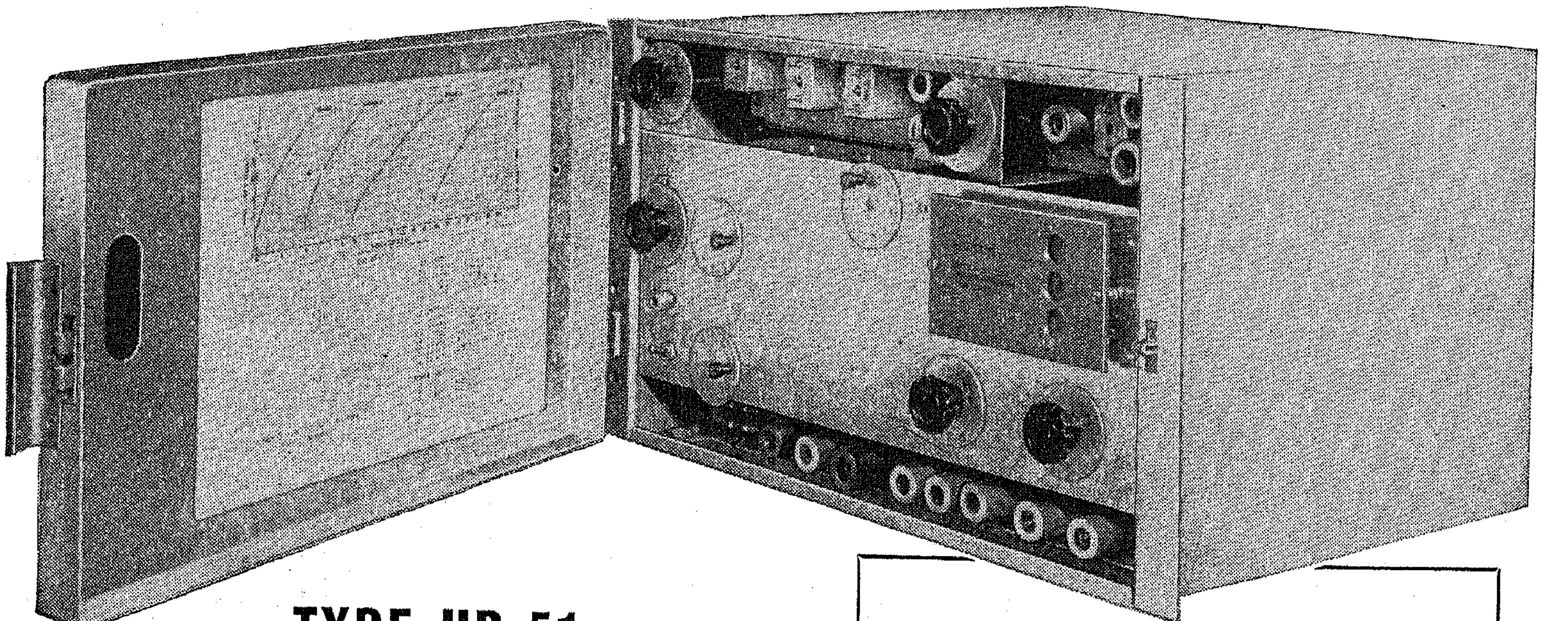
Assistant Secretary
F. C. HARRIS.

Editor-in-Chief
G. E. WILLIAMS, B.Sc.(Eng.), M.I.E.E.


ERIE "Silertex" Wirewound Resistors
TUBULAR MINIATURE SERIES 4-10 Watts 10-10,000 Ohms
TUBULAR POWER SERIES 10-50 Watts 10-40,000 Ohms up to 5 Sections
OVAL POWER SERIES 15-45 Watts 10-50000 Ohms up to 4 Sections
The excellence of Erie wirewound resistors is due, not only to the sound workmanship always associated with Erie products, but also to the coating of silicone bonded resin, known as "Silertex", which has all the advantages of vitreous enamel, plus the added advantage of a much lower firing temperature, and consequent freedom from surface crazing and, above all, from oxidation of the resistance wire.
That is why they are reliable above all others, and that is why you will find them used by those manufacturers who take pride in quality.
SERVICE COSTS MONEY – SO FIT "SILERTEX" TODAY
ERIE* Resistor Ltd
*Registered Trade Marks
Carlisle Road, The Hyde, London, N. W. 9., England. Telephone: COLindale 8011.
Factories: London and Great Yarmouth, England; Trenton, Ontario, Canada; Erie, Pa., and Holly Springs, Miss., U. S. A.


Life line of communication . . .
World wide radio-communication began with Marconi's Transatlantic messages in 1901. Since then Marconi research and development have been behind every major advance in technique. Marconi equipment today, operating at all frequencies, covers a very wide field of both long and short range radio/telegraph and radio/telephone requirements. Marconi VHF multi-channel equipment can provide for as many as 48 telephone channels and is largely superseding land line or cable routes on grounds of efficiency, economy, ease of installation and maintenance.
COMPLETE RADIO/TELEPHONE
AND RADIO/TELEGRAPH
SYSTEMS AND EQUIPMENT
MARCONI
COMPLETE COMMUNICATION SYSTEMS
Surveyed, Planned, Installed, Maintained
MARCONI'S WIRELESS TELEGRAPH COMPANY LIMITED · CHELMSFORD · ESSEX
LG I

Marconi Three Channel Telegraph/Telephone H.F. Receiver for remote control



TYPE HR 51

The Type HR 51 equipment is suitable for reception of telegraph or telephone signals on any one of three pre-set HF channels. It may be remotely controlled for channel selection and fine tuning from a distance of up to 10 miles. Control can be over the same wires as carry the AF signal output or a separate pair, provided the control circuit does not exceed 1000 ohms loop resistance. The receiver can be used to operate a recording unit such as the Marconi HU 11. Two may be connected for diversity reception, feeding a recording unit such as the Marconi HU 12.

Power supply components are housed in a compact bench mounting cabinet with the receiver. Access to all receiver controls and the valves is by a hinged door, which protects the controls from accidental interference. Lamps indicating the selected channel are visible through an aperture when the door is shut. Further access is by removable panels. The HR 51 is also available for rack mounting with associated equipment. The remote control unit, not shown here, is suitable for bench or rack mounting and requires connection to a mains supply point.

Over 80 countries now have Marconi equipped communications systems. Many of these are still giving trouble free service after more than twenty years in operation.

Lifeline of communication

MARCONI

COMPLETE COMMUNICATION SYSTEMS

Surveyed, planned, installed, maintained

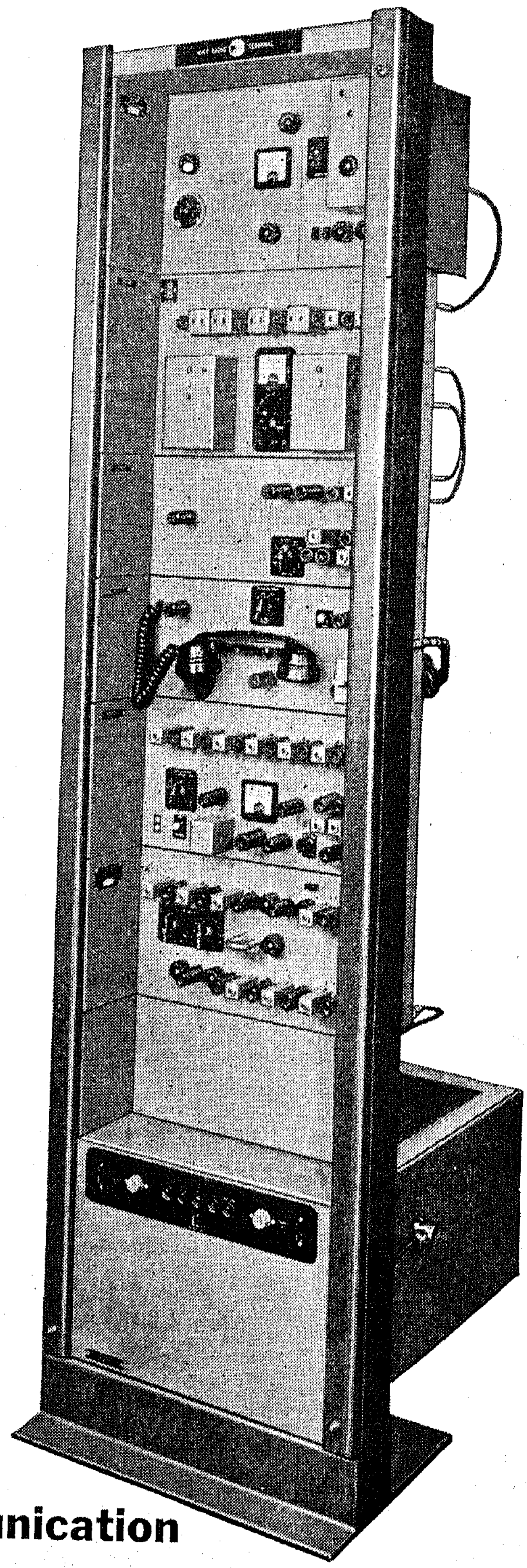
Partners in progress with The 'ENGLISH ELECTRIC' *Company Limited*

MARCONI'S WIRELESS TELEGRAPH CO. LTD., CHELMSFORD, ESSEX

LC 14

Marconi VHF Multi-Channel Equipment

TYPE HM 181



Multi-channel radio links are not only recognised economic alternatives to line and cable routes wherever the latter are costly because of intensive urban development or the wild nature of the terrain; they are frequently preferable in their own right. The type HM 181 equipment has been designed for comparatively simple schemes using two terminals working point-to-point or with a limited number of repeaters. It operates in the frequency range 150-200 **Mc/s**, employs frequency modulation and gives high performance with low distortion.

It provides the following facilities:—

- 8, 16 or 24 channels
- Repeaters with easy channel dropping facilities
- Unattended operation
- Engineers' order wire
- Ease of access for maintenance

Over 80 countries now have Marconi equipped telegraph and communication systems. Many of these are still giving trouble free service after more than twenty years in operation.

Lifeline of communication

MARCONI

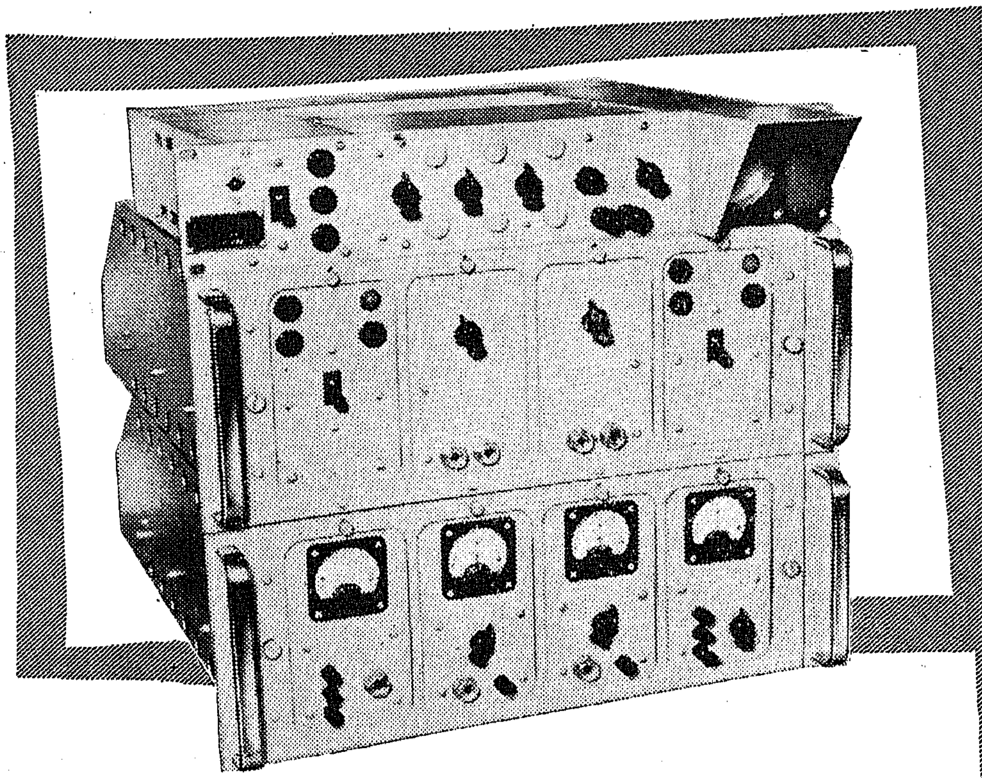
COMPLETE COMMUNICATION SYSTEMS

Surveyed, planned, installed, maintained

MARCONI'S WIRELESS TELEGRAPH CO., LTD., CHELMSFORD, ESSEX

Partners in progress with The 'ENGLISH ELECTRIC' *Company Ltd.*

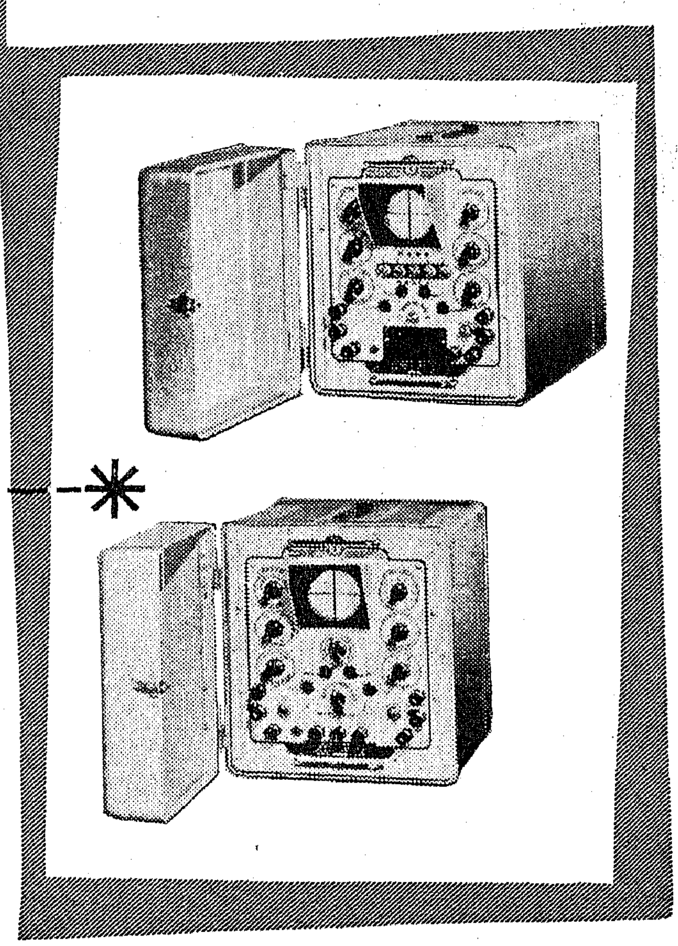
LC 7



✱ FREQUENCY SHIFT TELEGRAPH TERMINAL EQUIPMENT

Designed to work in conjunction with conventional receivers for the reception in dual diversity, of wide or narrow band frequency-shift and on/off, or reversed on/off, hand or automatic radio telegraph and teleprinter signals. Up to 85 db of rapid variation in input signal level can be accepted with frequency-shift working, and up to 35 db with on/off or reversed on/off, working. Keying speeds up to 200 bauds can normally be handled—this range can be extended if required. This versatile receiver is also suitable for use with the new 50 c/s Pilot Carrier frequency shift system.

A.T.E. Telegraph Equipment



TELEGRAPH DISTORTION MEASURING SETS

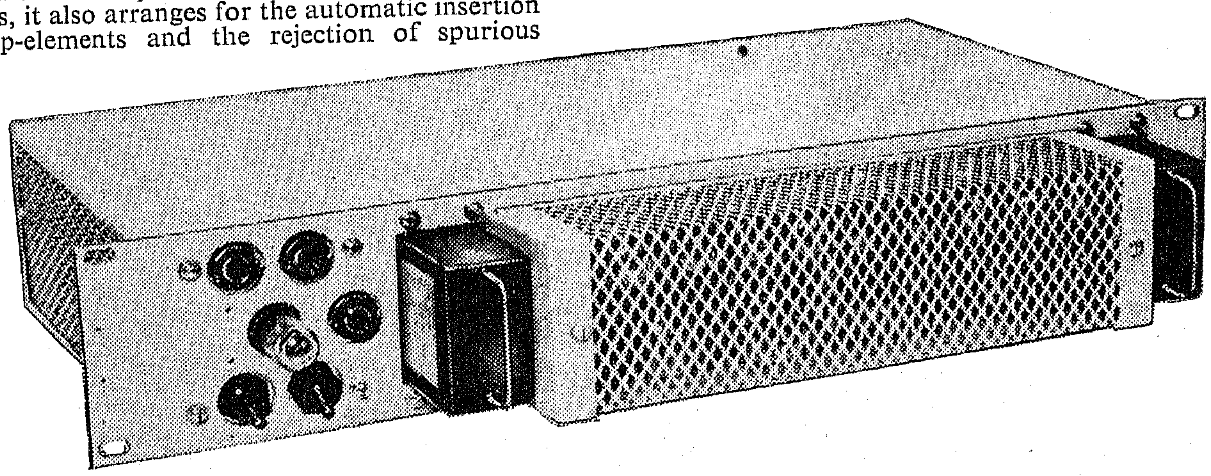
This equipment is available either in portable form or arranged for standard width rack mounting. There are two units each 18½″ x 11½″ x 13½″, both mains driven, either may be used independently for certain tests or both may be used in combination to cover a comprehensive series of tests. These tests, which need not interfere with normal transmission, cover transmission and reception. The transmitting unit can send perfect or distorted signals at any speed from 20 to 80 bauds or up to 200 bauds with modification. It can generate reversals and character repetitions and incorporates a 100 character test message sender. An additional feature of this unit is its use as a relay tester.

The receiver unit indicates the distortion on a *working circuit* without interrupting the service. Each element of a start-stop signal appears separately on the CRT which produces a spiral time base display. Adjustable speeds from 20 - 80 bauds or up to 200 bauds with modification.

Details of these and other telegraph equipments will gladly be sent on request.

✱ REGENERATIVE REPEATER

A mains operated, start-stop, five unit code equipment. Designed for use in both radio and line teleprinter circuits to regenerate and correct distorted signals, it also arranges for the automatic insertion of correct length stop-elements and the rejection of spurious signals.

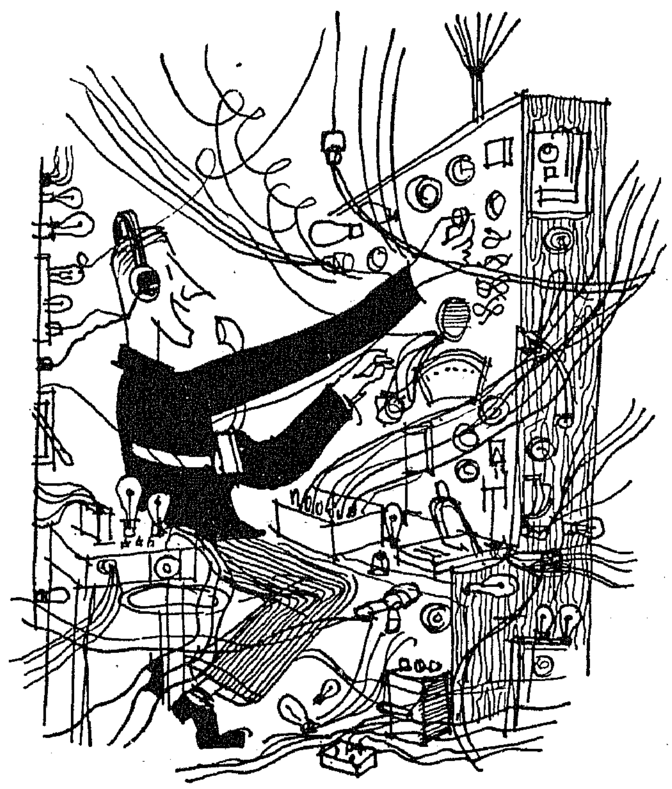


REGD. TRADE MARK

AUTOMATIC TELEPHONE & ELECTRIC COMPANY LTD.

(RADIO & TRANSMISSION DIVISION) *Strowger House, Arundel Street, London, W.C.2.*
Telephone: TEMple Bar 9262. *Telegrams: Strowgerex London.* Strowger Works, Liverpool 7.

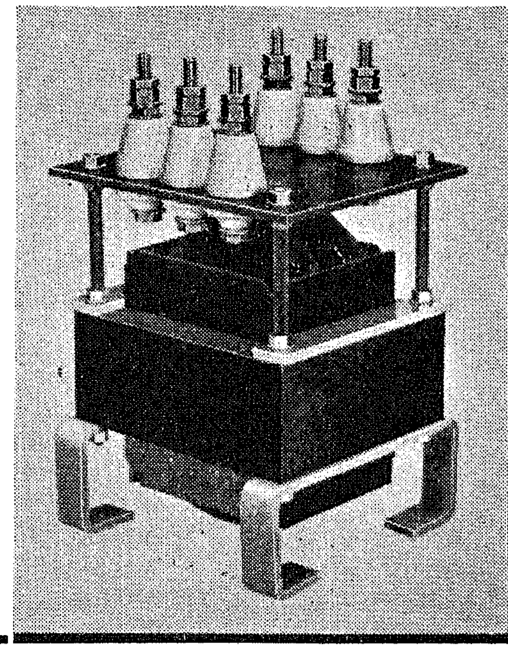
AT14551-BX107



no such thing as a difficult customer at

SAVAGE

This is the name usually associated with the *discerning* customer who always appreciates the quality and service by which we have gained our reputation.



SAVAGE TRANSFORMERS OF DEVIZES LIMITED

SAVAGE TRANSFORMERS LTD., NURSTEED RD., DEVIZES, WILTS. TEL. DEVIZES 932

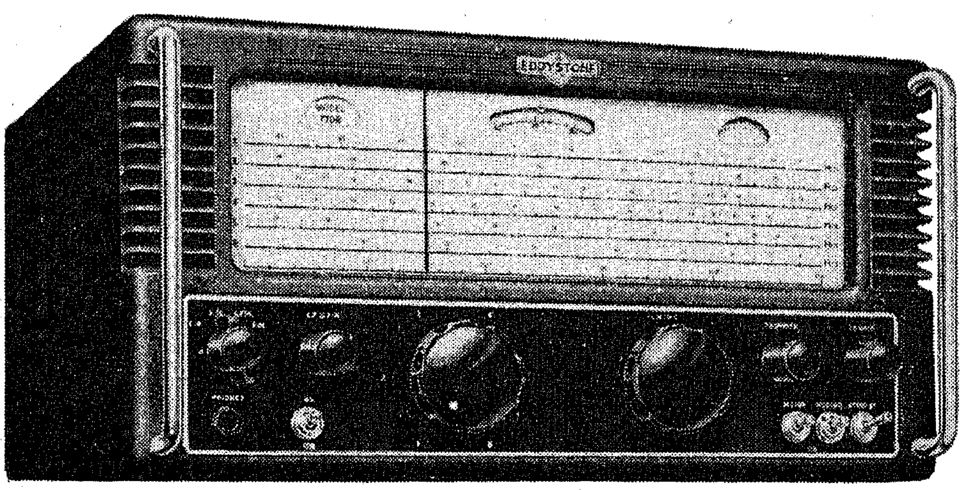
EDDYSTONE

V.H.F. COMMUNICATIONS RECEIVERS

MODEL 770R – 19 Mc/s. to 165 Mc/s. CONTINUOUS

MODEL 770U – 150 Mc/s. to 500 Mc/s. CONTINUOUS

FOR MONITORING, FIELD TESTS, LABORATORY PURPOSES ETC.

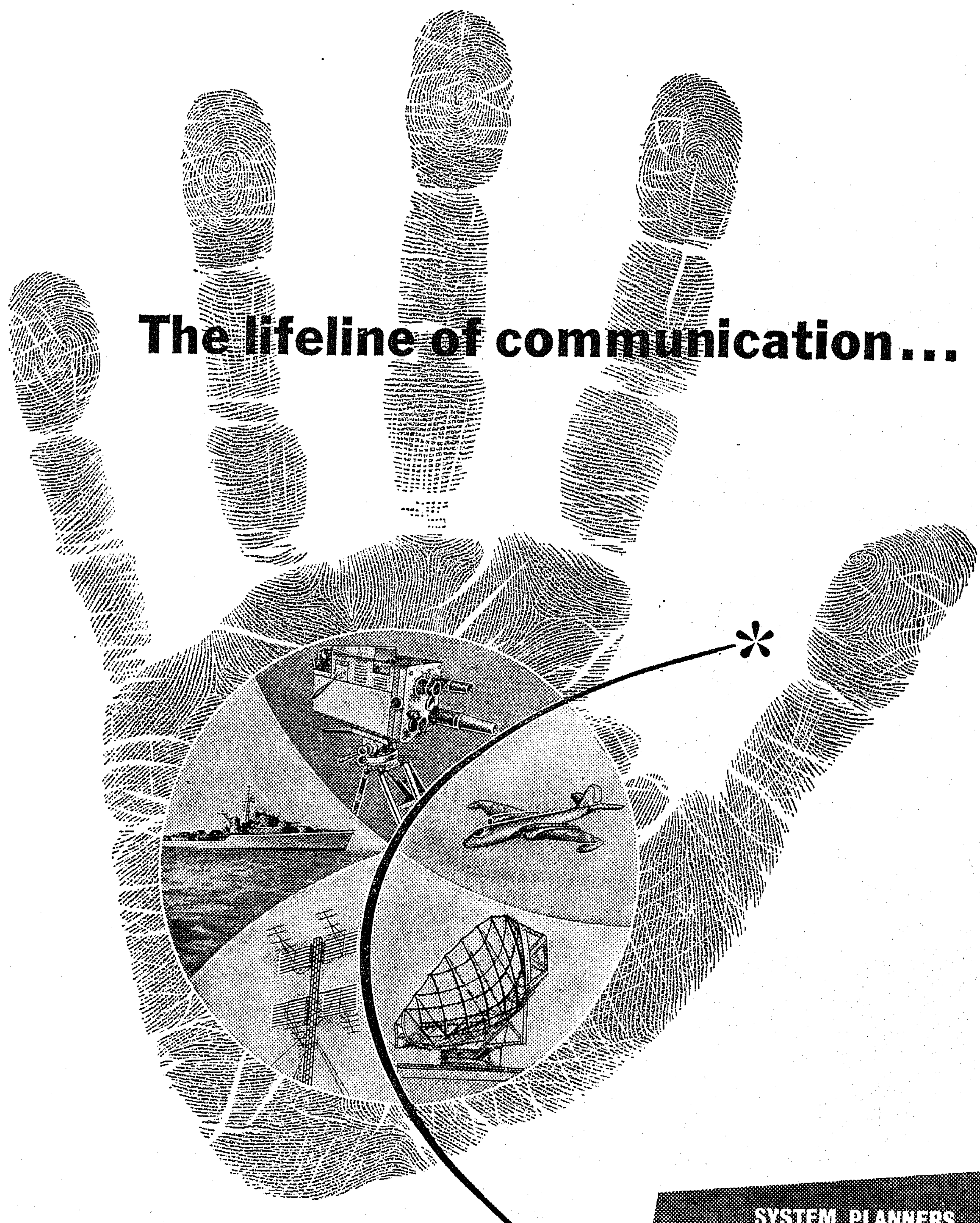


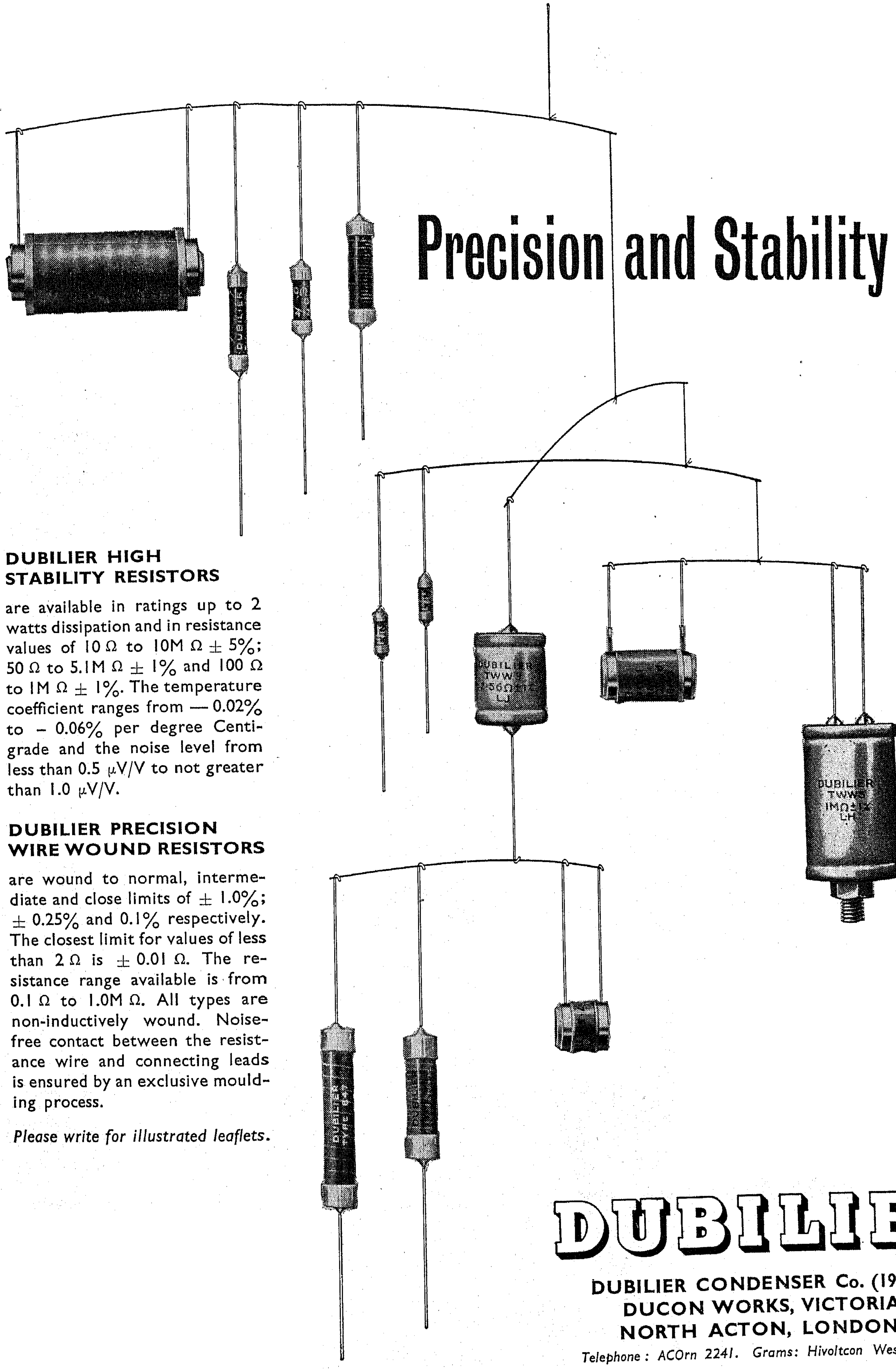
- Highly efficient signal frequency circuits.
- Substantial diecast rotary coil turret.
- Excellent frequency stability and selectivity.
- Accurate re-setting and ease of handling.
- High sensitivity and excellent signal-to-noise ratio.
- For AM and FM.
- Robust construction and outstanding reliability.
- "S" Meter. Noise Limiter.
- Preferred type valves.
- Finest workmanship.

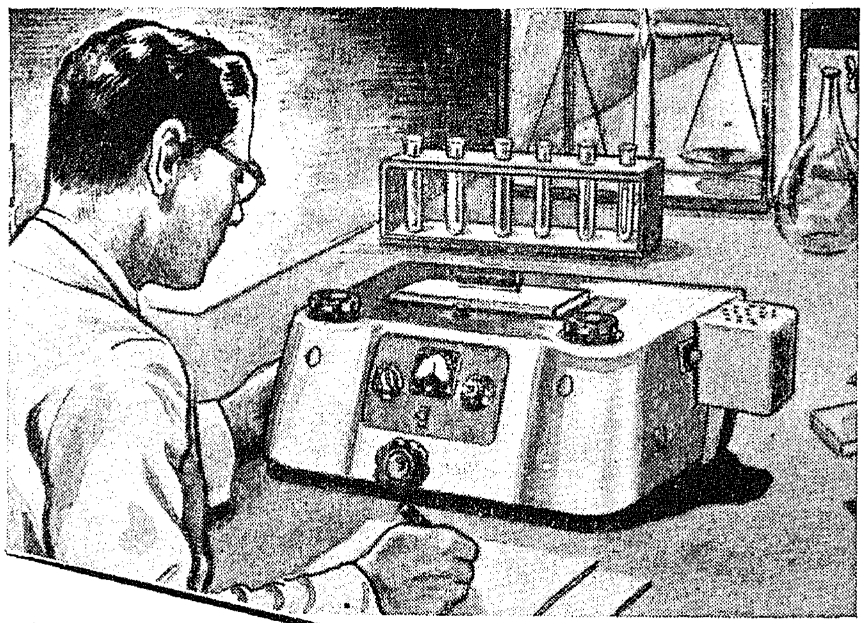
- A. C. operation. 110-250 volts. 40-60 cycles.
- Dimensions 16¾″ x 15″ x 8¾″.
- Weight 60 lbs.

PLEASE WRITE FOR FULL SPECIFICATION TO THE MANUFACTURERS:

STRATTON & CO. LTD., ALVECHURCH ROAD, BIRMINGHAM, 31


The lifeline of communication...





the Research Chemist relies upon

CONSTANT VOLTAGE....

In problems of analysis to-day, the chemist is aided to an increasing degree by electronics. However, spectroscopic techniques and apparatus depend for their accuracy on a reliable constant voltage supply, and here Advance Constant Voltage Transformers provide the answer. Purely automatic, and with no thermionic valves or moving parts, they can be relied upon to provide the stabilized A.C. voltage so necessary in the laboratory.

Full details given in folder A28 gladly sent on request.

(Illustration of spectrophotometer by courtesy of Unicam Instruments.)

Advance

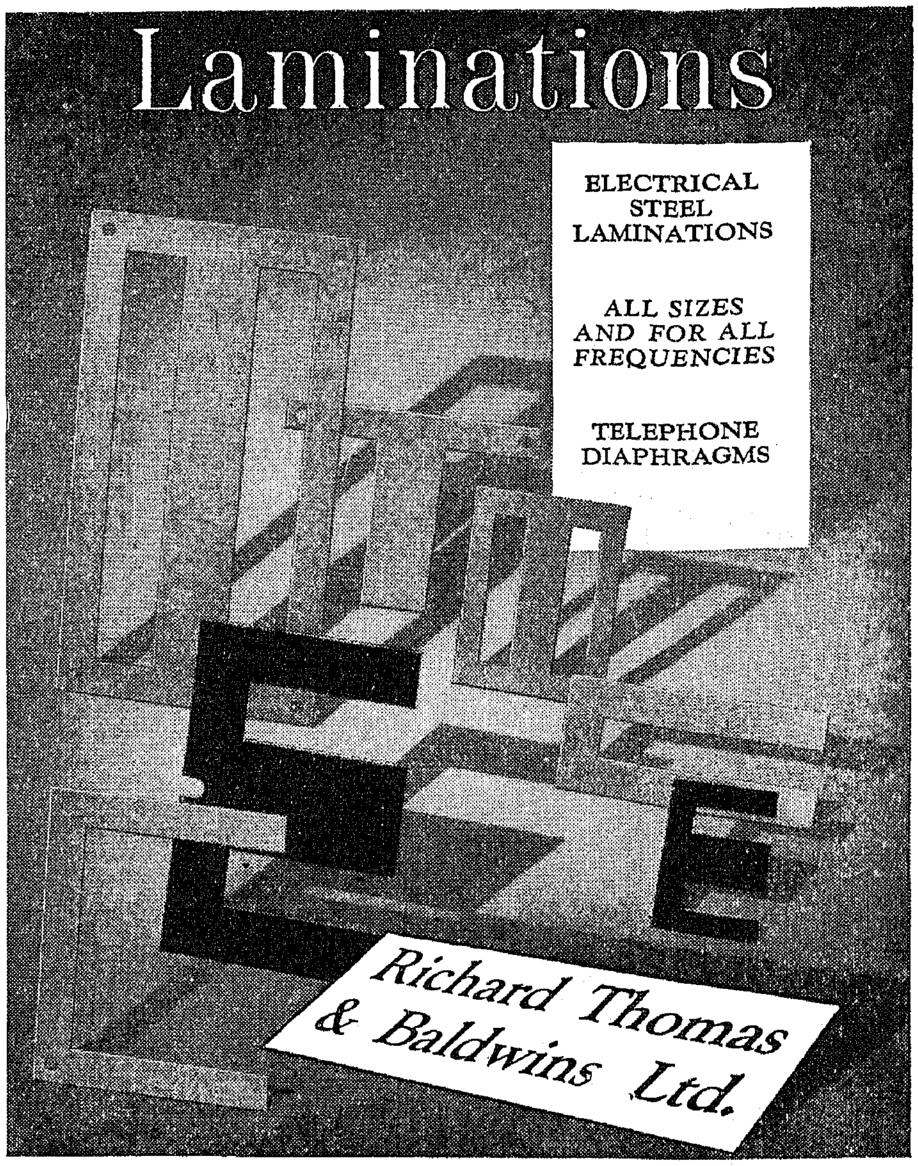
CONSTANT VOLTAGE TRANSFORMERS

CVT

ADVANCE COMPONENTS LTD.

MARLOWE ROAD, WALTHAMSTOW, LONDON, E.17

Telephone: LARkswood 4366/7/8



A MEMBER OF THE RTSC GROUP
COOKLEY WORKS, BRIERLEY HILL, STAFFS.
Head Office: RTSC HOUSE, PARK ST, LONDON, W.1

WHY IT PAYS TO USE Ersin Multicore Solder

Radio Manufacturers all over the world prefer to use Ersin Multicore Solder in their factories and workshops. Even in the U.S.A., where more cored solder is produced than anywhere else in the world, many leading firms insist on British made Ersin Multicore. Below are some of the reasons why Ersin Multicore has attained such world-wide popularity.

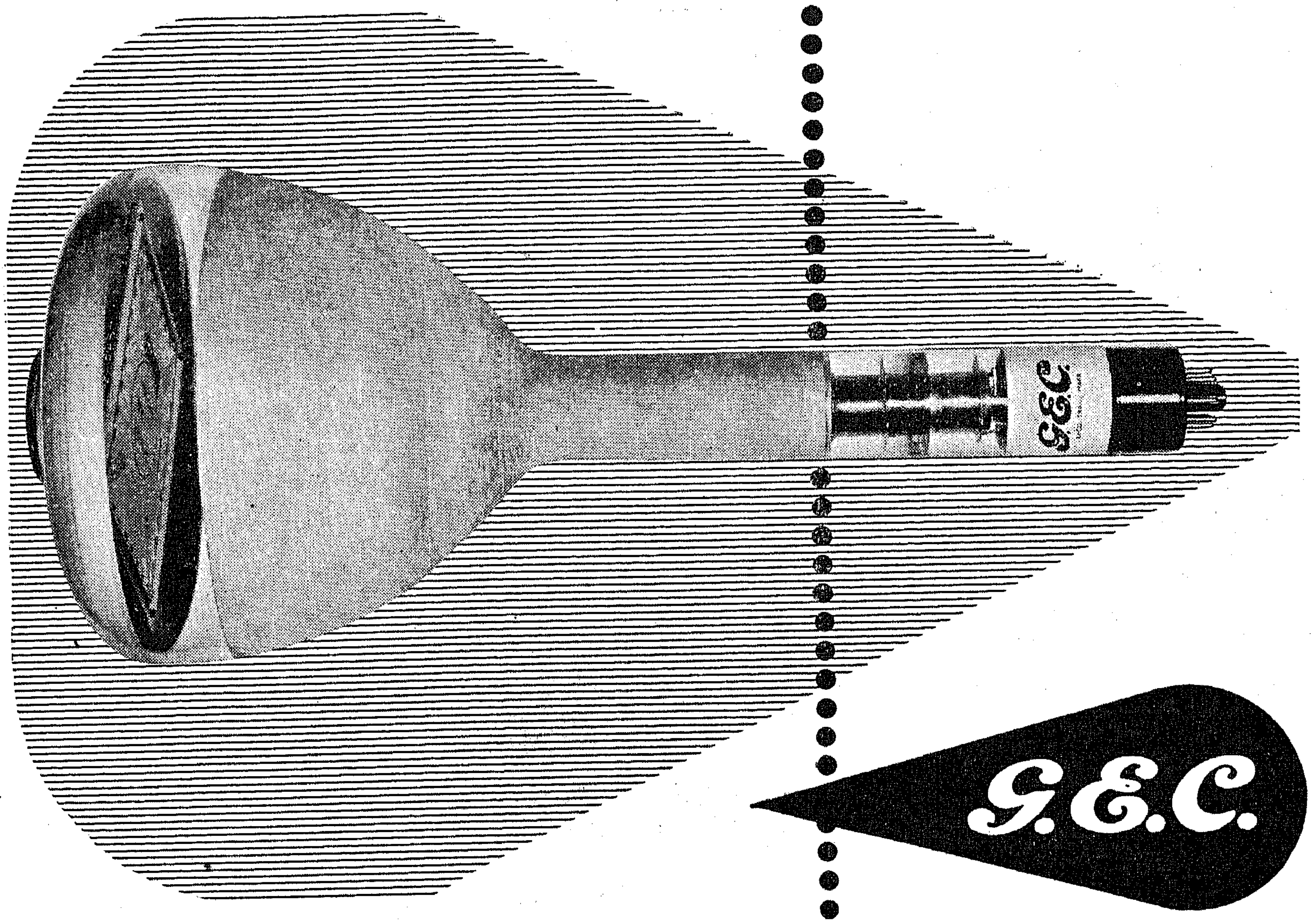
- Ersin Multicore is the only solder containing 5 cores of Ersin Flux, a high grade rosin which has been subjected to a complex chemical process to increase its fluxing action, whilst still retaining the non-corrosive properties. Ersin Flux makes precision soldering quicker and more economical—it not only prevents oxidation during soldering but actually cleans the surface to be soldered, removing any oxide from the metal.
- Five cores of flux ensure flux continuity throughout the length of the solder wire—there are no lengths without flux.
- The correct proportions of flux to solder are always assured—no extra flux is required. Five cores of flux provide thinner solder walls, giving instantaneous melting.
- Soldered joints made with Ersin Flux do not corrode even after prolonged exposure to any degree of humidity.
- Only the finest virgin tin and lead are used in the manufacture of Ersin Multicore.

FOR FACTORY USE. The economies effected by using Ersin Multicore Solder play an important part in cutting production costs and keeping down the price of equipment. You get more joints per lb. of Ersin Multicore—there is no waste. Soldering with Ersin Multicore is quicker too and every joint is a perfect electrical connection. Ersin Multicore Solder is made as standard for factory use in 6 alloys and 9 gauges, and is supplied on nominal 1 lb. and 7 lb. reels. Other Alloys and Gauges can be supplied to special order. Bulk prices on application.

TECHNICAL INFORMATION. Electrical engineers and technicians are invited to write for comprehensive technical literature about Ersin Multicore Solder containing useful tables of melting points, etc., and samples of alloys.

MULTICORE SOLDERS LTD.

Multicore Works, Hemel Hempstead, Herts. Boxmoor 3636



G.E.C.

1693 HKM

MONOSCOPE

The G.E.C. Monoscope is a secondary emission type of picture signal generator tube.

It offers the simplest means of providing a video test signal, which is obtained by scanning a fixed pattern reproduced on a target plate within the tube.

The derived picture may consist of either line or half tone information or both in combination, with a maximum vertical definition of 800 lines. The picture plate dimensions are 98 × 74 mm giving the standard aspect ratio of 4 × 3.

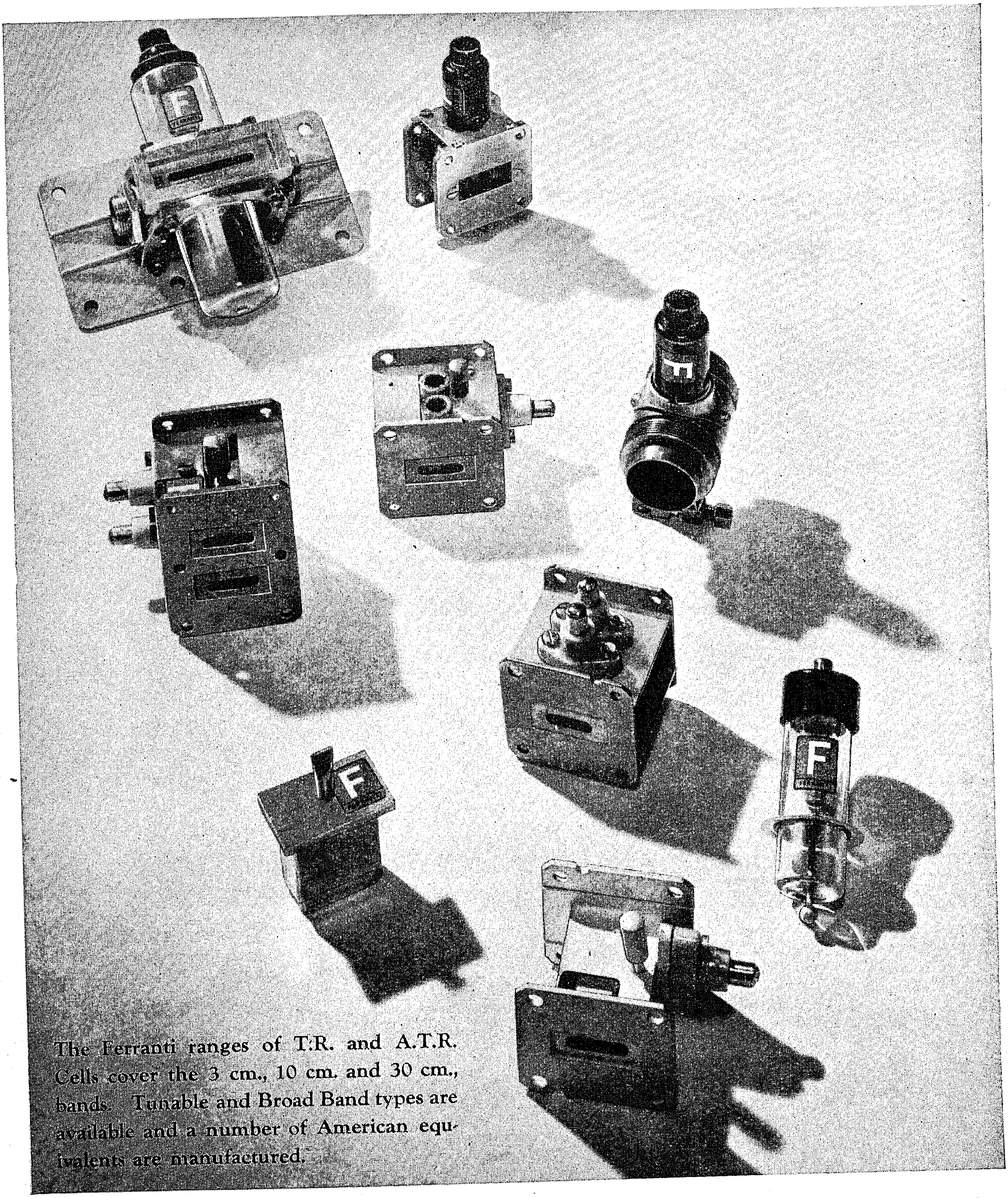
The standard pattern is test card "C", but alternative patterns can normally be supplied to meet special requirements.

TYPICAL OPERATION

V_{a1}		1.25	kV
V_{a3}		5	kV
V_g	for cut-off	—60	V
V_t	to A_3	—120	V
I_b		5 to 10	μA
I_{sig}	(pk to pk)	$I_b/4$ to $I_b/2$	μA
R_L		2 to 5	kΩ

THE GENERAL ELECTRIC CO. LTD., MAGNET HOUSE, KINGSWAY, LONDON, W.C.2

FERRANTI T.R. and A.T.R. CELLS for RADAR EQUIPMENT



FERRANTI

FERRANTI LTD. CREWE TOLL, FERRY ROAD, EDINBURGH 5

London Office: KERN HOUSE, 36, KINGSWAY, W.C.2.

THE PROFESSIONAL ENGINEERS APPOINTMENTS BUREAU

Incorporated under Limited Guarantee, 1947
(Licensed annually by the London County Council)
9 VICTORIA STREET, LONDON, S.W.1

The Principal Objectives of the Bureau are:

1. *To Aid Employers.*

To receive inquiries from employers seeking the services of qualified professional engineers and to advise such employers, if required, as to the qualifications desirable for the particular vacancy.

To submit to employers particulars of persons registered with the Bureau, the qualifications of whom seem to fit them for the appointments.

2. *To aid Members of the Institutions of Civil, Mechanical and Electrical Engineers.*

To receive applications for registration for employment from engineers, who by reason of their engineering qualifications are Corporate or Non-Corporate Members of the Institution of Civil Engineers, The Institution of Mechanical Engineers, or The Institution of Electrical Engineers (excluding Associates of The Institution of Civil Engineers); and to charge and receive such registration fees and appointment fees as may from time to time be determined by the Bureau.

To give advice generally in matters relating to the employment of professional engineers.

EMPLOYERS

Employers of professional engineers who have vacancies to be filled, are invited to notify such vacancies to the Bureau and to furnish the following particulars:

(1) Title of post with description of duties involved.
(2) Age range. This should be made as wide as possible.
(3) Salary offered and supplementary remuneration, if any.
(4) Professional, technical and/or special qualifications, e.g. languages.
(5) Location of work.
(6) Could a successful candidate be assisted in finding suitable living accommodation locally?
(7) Does the post involve control of staff? If so, to what extent?
(8) Any other information which will assist in the selection of candidates.

Notifications of vacancies in Commonwealth countries are particularly welcome.

The Bureau deals with all classes of appointments ranging from practical training for Students to executive or administrative posts for senior Corporate Members.

MEMBERS OF THE THREE INSTITUTIONS

Members of the three Institutions who wish to register with the Bureau for employment may obtain the necessary forms on application to the Registrar. The grade of membership should be stated, and a stamped addressed foolscap envelope should be enclosed.

The Professional Engineers Appointments Bureau receives a number of applications for assistance from young graduate engineers, Civil, Mechanical and Electrical, who require practical experience as trainees. The Bureau would therefore be grateful to receive from such firms advice of such vacancies as they occur, or when annual intakes are being considered.

A remarkable NEW head telephone set by Standard

Here's the most interesting telephone news for years ! A new, lightweight, one-piece head telephone set moulded in nylon plastic and weighing only 4 oz !

Available in black or ivory finish, this revolutionary instrument, designed, developed and manufactured by Standard Telephones & Cables Ltd., for telephone switchboard operators, introduces an entirely new concept in comfort, elegance, economy and operational efficiency.

No other operator's head telephone set available to-day has all these additional advantages :

Self-contained Nylon moulding incorporating high-quality, super sensitive transmitter and receiver. *

Weight only 4 oz (compare this with the 15 oz weight of the conventional head and breast type operators' set). *

Attractive appearance. *

Only one instrument cord required. *

Finely balanced to ensure comfort and stability. *

Mechanical shock absorber to avoid breakage at transmitter horn junction. *

Adjustable head grip with resilient band. *

Moulded frame for operator's identification label. *

Another advance by Standard

Ask for brochure T.D.16 which fully describes the

4406 Telephone Head Set

Standard Telephones and Cables Limited

TELEPHONE DIVISION

OAKLEIGH ROAD · NEW SOUTHGATE · LONDON, N.11

Registered Office: Connaught House, Aldwych, London, W.C.2

THE PROCEEDINGS OF THE INSTITUTION OF ELECTRICAL ENGINEERS

TEN-YEAR INDEX

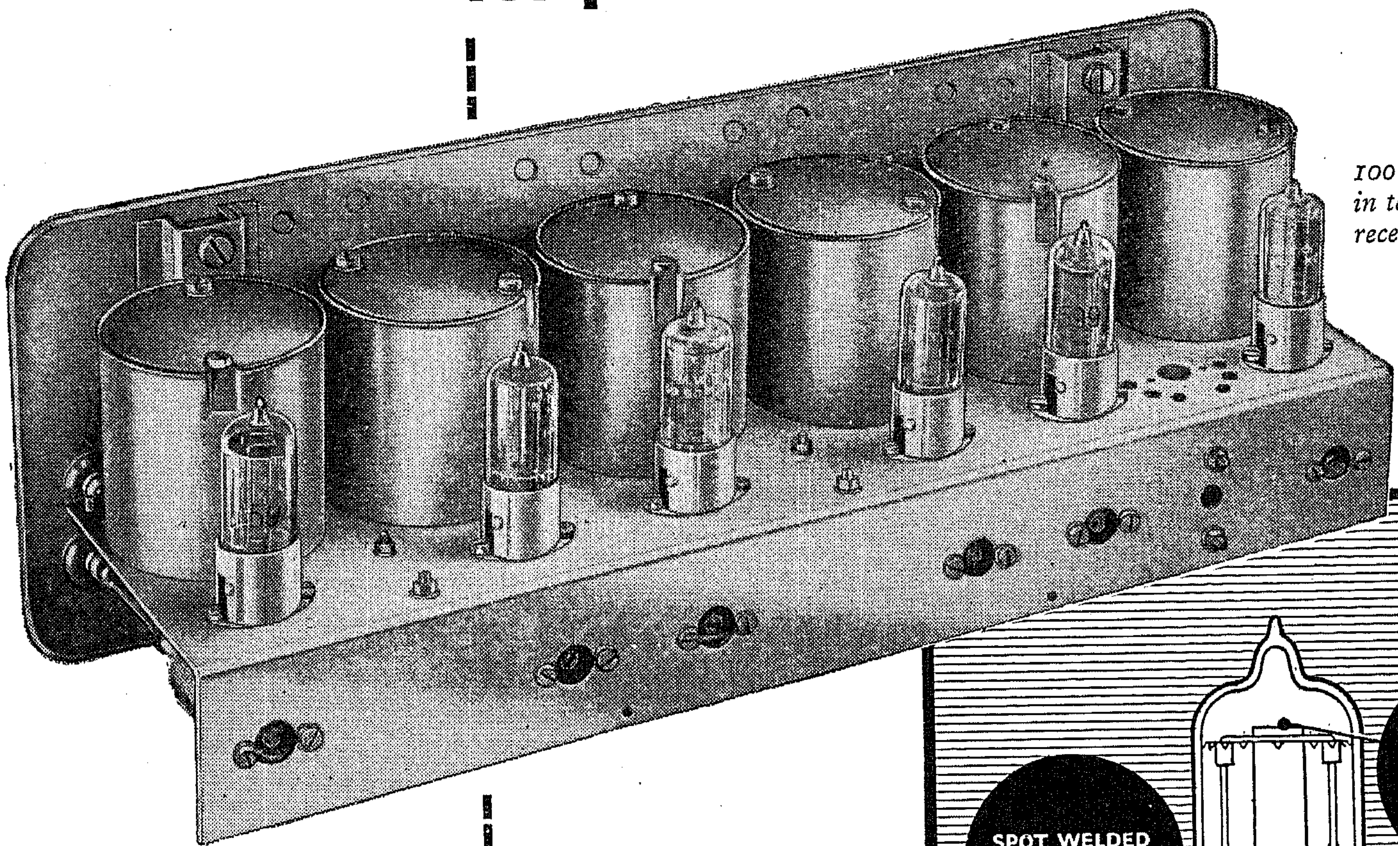
1942—1951

A TEN-YEAR INDEX to the *Journal of The Institution of Electrical Engineers* for the years 1942–48 and the *Proceedings* 1949–51 (vols. 89–98 can be obtained on application to the Secretary.

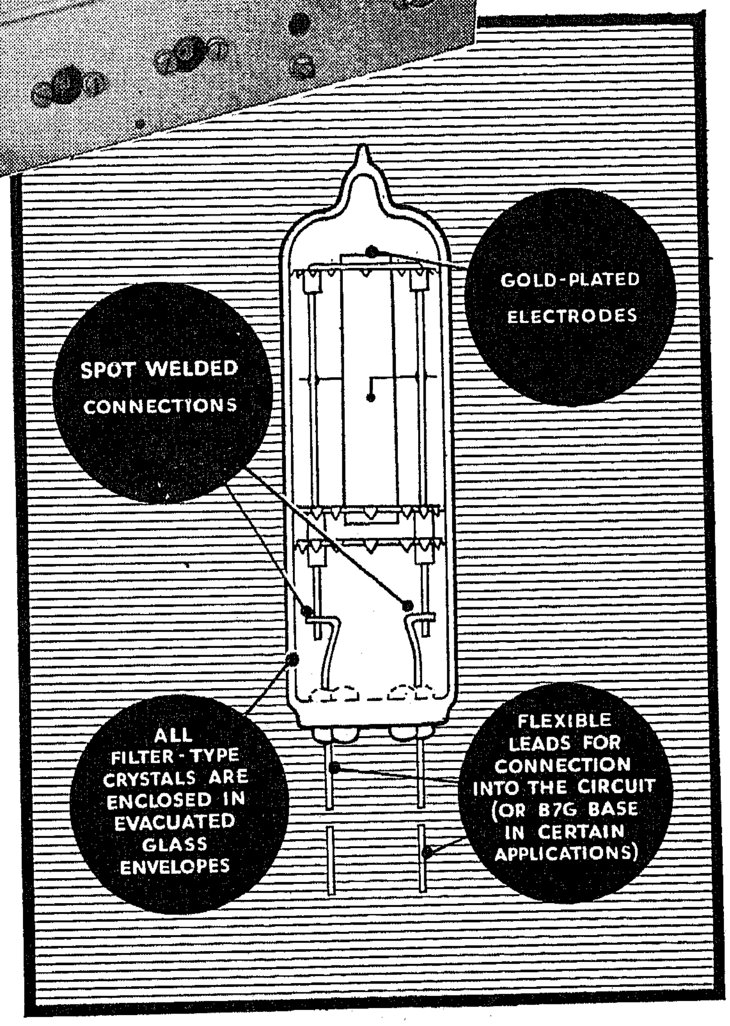
The published price is £1 5s. od. (post free), but any member of The Institution may have a copy at the reduced price of £1 (post free).

MARCONI CRYSTALS

for precision filters 40-700 Kc/s



100 Kc/s crystals in telegraph receiver filter



Available in standard ranges or for customer-designed units

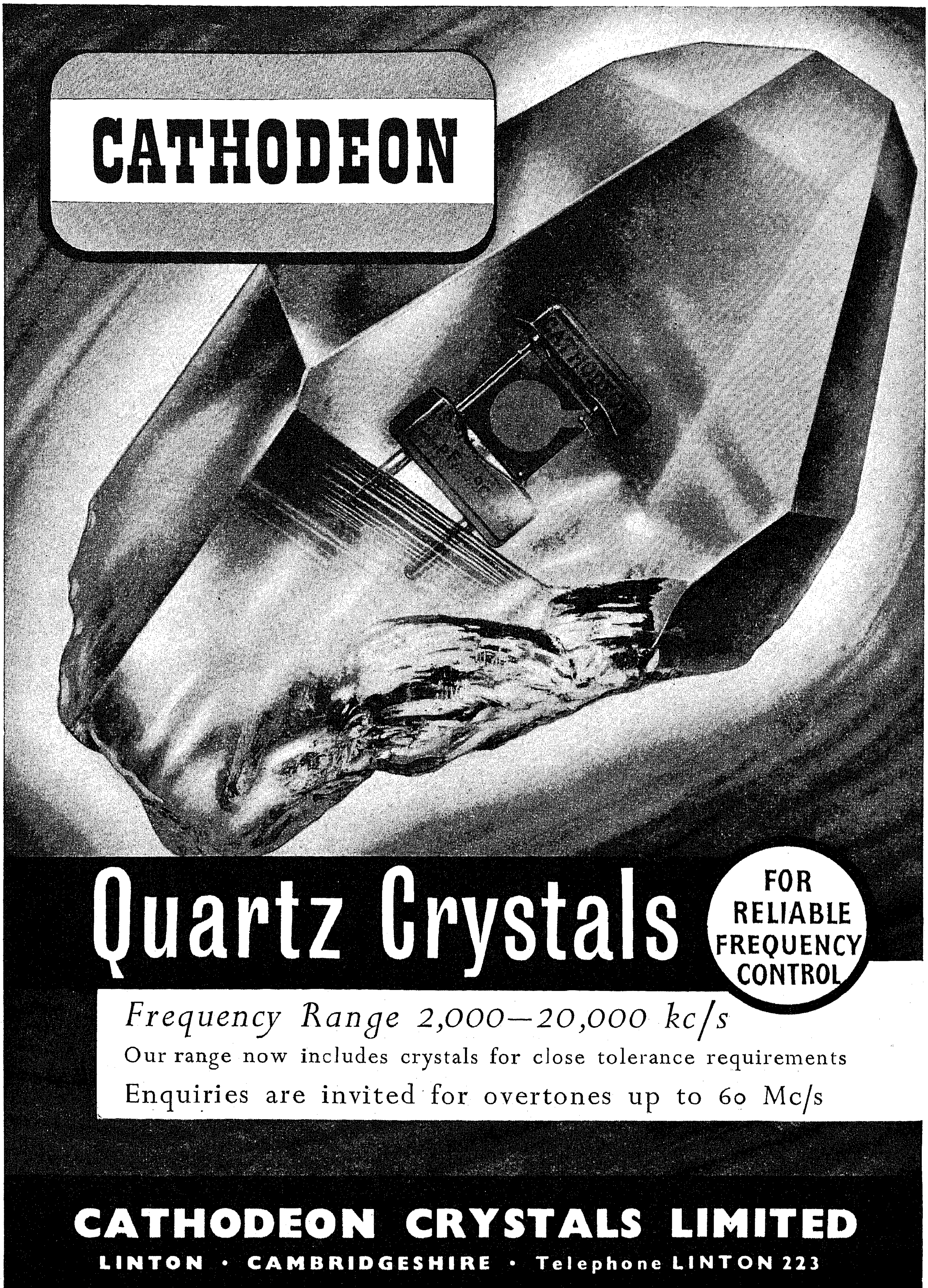
The experience gained in manufacturing quartz crystals to the stringent requirements of our own apparatus enables us to offer a comprehensive range of crystals for precision filters. Years of intensive research and development work in this field ensure the reliability and quality of this Marconi product.

Lifeline of communication

MARCONI

Partners in progress with the 'ENGLISH ELECTRIC' *Company Limited*

MARCONI'S WIRELESS TELEGRAPH CO., LTD., CHELMSFORD, ESSEX. Tel: Chelmsford 3221 CR 3

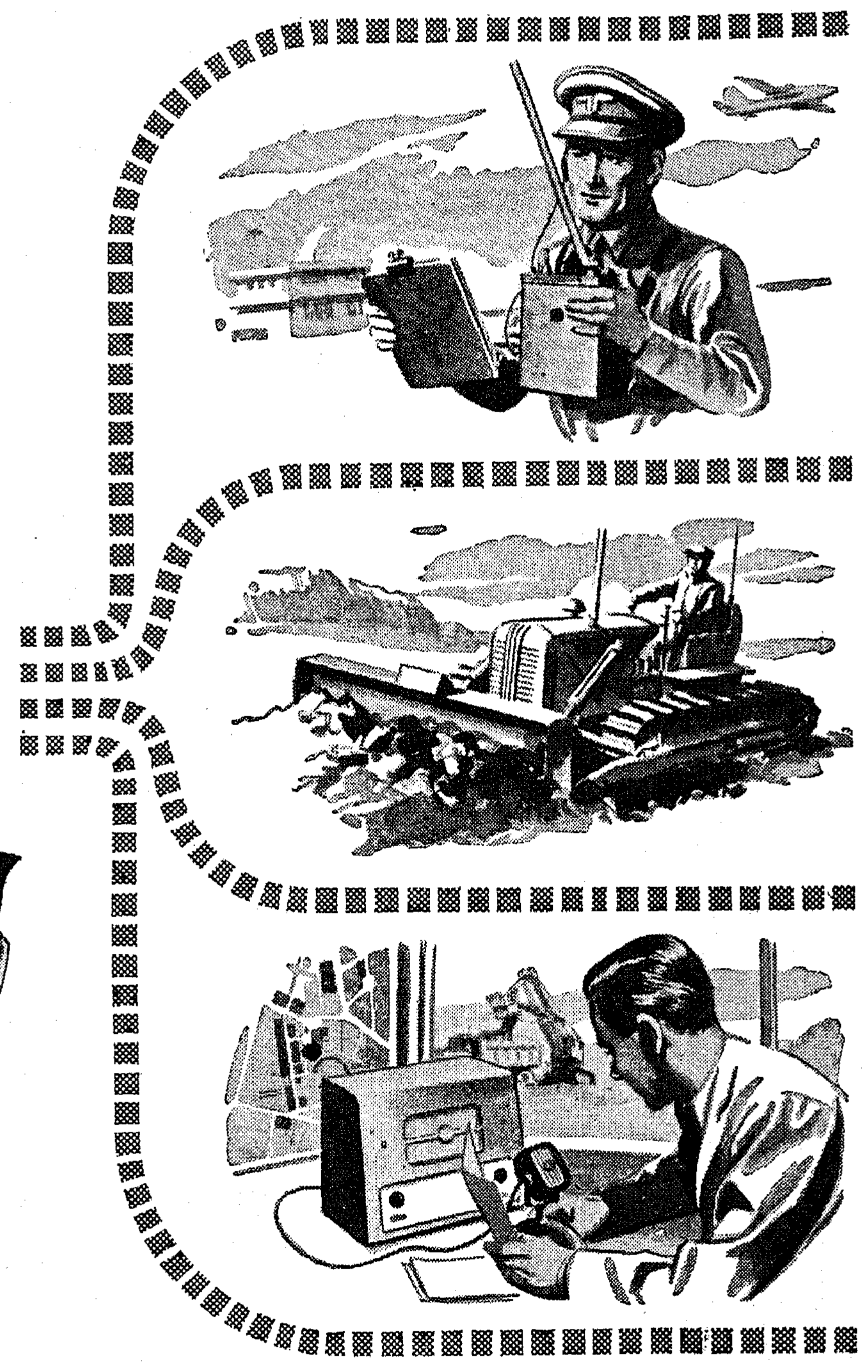


CATHODEON
Quartz Crystals
FOR RELIABLE FREQUENCY CONTROL
Frequency Range 2,000—20,000 kc/s
Our range now includes crystals for close tolerance requirements
Enquiries are invited for overtones up to 60 Mc/s
CATHODEON CRYSTALS LIMITED
LINTON · CAMBRIDGESHIRE · Telephone LINTON 223

PYE WALKIEPHONE

PORTABLE RADIO TELEPHONE FOR PERSONAL CONTACT WITH KEY MEN

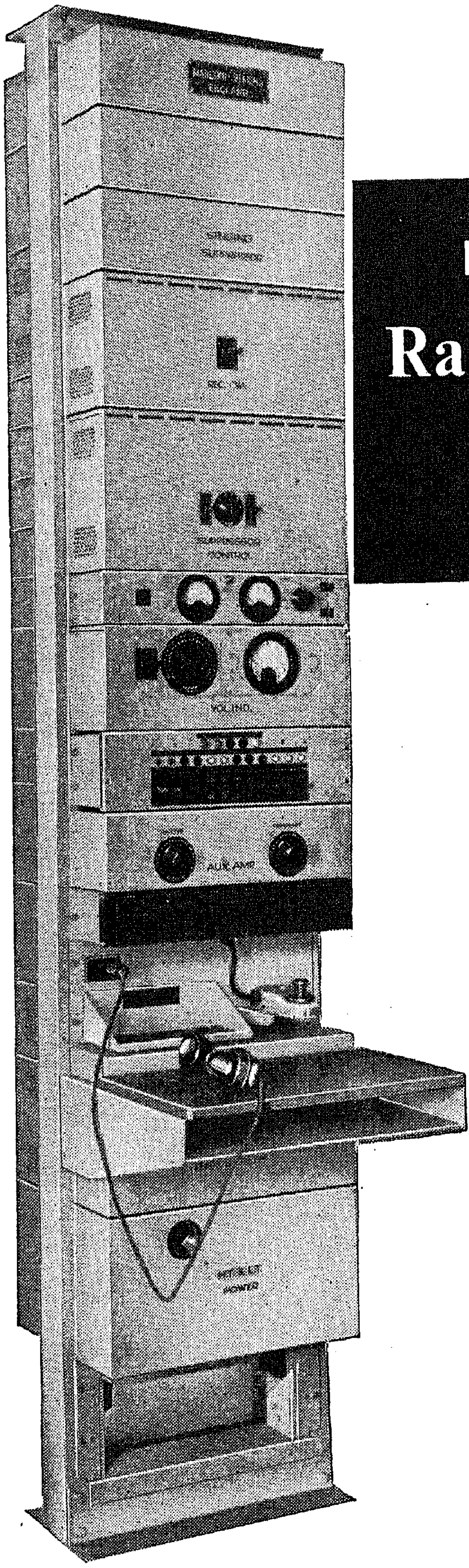
Policemen, Firemen, Foremen, Airfield Supervisors and many other key men are more effective when in constant radio contact with their headquarters. The Pye Walkiephone, which weighs only $10\frac{1}{2}$ lb., is ideal for this purpose. An outstanding tribute to the Walkiephone was paid by Sir John Hunt, who praised the part it played in the victorious assault on Everest.



PYE
Telecommunications
CAMBRIDGE ENGLAND

Pye (New Zealand) Ltd. Auckland C.I., New Zealand	Pye Canada Ltd. Ajax, Canada	Pye-Electronic Pty., Ltd. Melbourne, Australia	Pye (Ireland), Ltd. Dublin, Eire
Pye Radio & Television (Pty.) Ltd. Johannesburg South Africa	Pye Limited Mexico City	Pye Limited Tucuman 829 Buenos Aires	Pye Corporation of America 270, Park Avenue New York

PYE LIMITED • • CAMBRIDGE • • ENGLAND
TEVERSHAM 311



THE PROCEEDINGS OF THE INSTITUTION OF ELECTRICAL ENGINEERS

EDITED UNDER THE SUPERINTENDENCE OF W. K. BRASHER, C.B.E., M.A., M.I.E.E., SECRETARY

VOL. 102. PART B. No. 6. NOVEMBER 1955

Paper No. 1925 R
Nov. 1955

621.396.2 : 621.391

AN INTRODUCTION TO SOME TECHNICAL FACTORS AFFECTING POINT-TO-POINT RADIOCOMMUNICATION SYSTEMS

By F. J. M. LAVER, B.Sc., Member.

(*The paper was first received 2nd May, and in revised form 21st July, 1955.*)

SUMMARY

Technical factors affecting point-to-point radiocommunication systems are briefly reviewed, as an introduction to the very extensive literature of the subject.

(1) INTRODUCTION

Point-to-point radiocommunication systems convey information between two fixed points and are judged by their fitness for this purpose. Information is a somewhat abstract entity, but it appears for communication as words, sounds or images, which are presented to the radio system in electrical form as telegraph, telephone, facsimile or television signals of standard types. The radio system's function is to convey these signals without significant corruption to a distant point, using freely propagated electromagnetic waves. Radio engineers are thus concerned with four main topics: information, signalling systems, radio equipment and radio-wave propagation.

(2) THE TRANSMISSION OF INFORMATION

Many of the theoretical results of information and communication theory[1–6] have long been familiar to practical engineers, although in a qualitative and intuitive way, and the main achievement has been to present a coherent and quantitative account of the nature and transmission of information which enables precise comparisons to be made between different communication systems. Communication theory is a difficult subject mathematically, but a brief consideration of the measurement of information is given below.

The function of a communication system is to reproduce at the distant end a series of symbols selected at the sending end. Whether the chosen symbols have any meaning does not enter into the estimate of the "information" that they convey. The important facts are that each symbol is chosen from a set of possible symbols, and that a certain degree of difficulty attaches to each possible choice. The greater the number of possible symbols the greater is the information conveyed by the choice of any one of them. The information conveyed by a symbol can be defined as the number of binary divisions needed to select it, and any selection from N symbols can be made by $\log_2 N$ binary choices. This number is the number of digits when N is expressed on the binary scale, and the information measure $\log_2 N$ is expressed in binary digits—"bits" for short. The measure can be extended to cover the choice of symbols from a set having different known probabilities of occurrence, and if p_i is the frequency with which symbol i occurs on the average in a statistically stable set of selections, the information per symbol is

$$H = -\sum_{1}^{N} p_i \log_2 p_i \quad \text{bits per symbol} \qquad (1)$$

The negative sign is used to make H positive, for p_i will be less than unity and $\log_2 p_i$ will thus be negative. When the probabilities p_i are all equal, H reaches its maximum value of $\log_2 N$ bits.

For a channel with a restricted number of signalling conditions, e.g. on-off keying, there is a maximum rate of flow of information which is called its capacity, say C bits/sec. When the channel is free from noise it can transmit symbols at an average rate of C/H symbols/sec; however, when the channel is noisy the original message cannot be reproduced with certainty by any operation on the received signal. Even so, it is possible to assign a definite capacity to a noisy channel, and by suitable coding of the signals it is theoretically possible to convey information at any rate up to this capacity with as small a frequency of errors as desired. It is generally impracticable to determine the ideal code for a particular channel, but an approach can be made by choosing a signal that remains more like the original than any other reasonable signal when it is affected by noise. This usually involves some redundancy. Redundancy is inherent in messages in plain language; in literary English, for instance, correlations between letters due to spelling, between words due to syntax and between sentences due to unity of thought, reduce its information per letter from $\log_2 26 = 4\cdot7$ bits/letter to about 1 bit/letter, i.e. its redundancy approaches 80%.[4]

In telephone and television systems the signals are continuously variable quantities theoretically capable of an infinite number of possible values, corresponding to an infinite amount of information. In practice the number of possible values is

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
This is an "integrating" paper. Members are invited to submit papers in this category, giving the full perspective of the developments leading to the present practice in a particular part of one of the branches of electrical science.
Mr. Laver is in the Post Office Engineering Department (W.P. Branch).

limited by noise and channel bandwidth. The sampling theorem states that a signal contained in a frequency band of width f_w c/s can be determined, i.e. reconstructed, by giving its values at a series of sampling points spaced at intervals of $1/(2f_w)$ sec, so that a continuous signal lasting T sec and confined in f_w c/s bandwidth can be represented by $2Tf_w$ numbers. Noise causes adjacent signal levels to overlap, and the number of distinguishable signals is approximately

$$[(P + N)/N]^{Tf_w} \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

where P and N are the average signal and noise powers respectively. The maximum amount of information that can be passed through this channel in time T is then

$$Tf_w \log_2 (1 + P/N) \text{ bits} \quad . \quad . \quad . \quad . \quad . \quad (3)$$

and its capacity is

$$C = f_w \log_2 (1 + P/N) \text{ bits/sec} \quad . \quad . \quad . \quad . \quad . \quad (4)$$

Eqn. (4) applies for signals disturbed by noise having a uniform spectrum, and its different terms can be varied to achieve a given capacity; for example, increasing the channel bandwidth allows the transmitter power to be reduced.

Practical systems using codes of reasonable complexity[7] fall considerably short of this ideal, which is most closely approached by pulse code modulation for high signal/noise ratios and pulse position modulation for low signal/noise ratios, although even these systems require some 3–5 dB more power than the theoretical minimum.[8]

(3) THE TRANSMISSION OF SIGNALS

In transmission over point-to-point radio systems signals are open to corruption by noise, distortion and interference from other radio signals. Some disturbance of the signal is unavoidable, and economic design aims only at reducing it to tolerable proportions. What is tolerable depends on the use made of the received signals. Higher standards of telegraph transmission are needed for messages in cipher than for those in plain language, and whereas for commercial telephony the main criterion is intelligibility, broadcasting adds further requirements. Tolerances for noise, distortion and interference have been determined by operating experience or subjective tests for some signals and some sources of disturbance, but not for all.

(3.1) Noise

(3.1.1) Sources of Noise.

Some sources of noise are peculiar to radio systems, e.g. atmospherics and extraterrestrial noise.[9] Atmospherics are a major factor at low radio frequencies, but above about 30 Mc/s their effect is generally lower than that of the noise generated in the radio receiver. This is partly because atmospheric pulses are relatively long, with an energy distribution which decreases at the higher frequencies, and partly because the characteristics of radio-wave propagation limit the reception of the higher-frequency atmospherics from distant storms. Thunderstorms are most frequent in tropical regions, and atmospheric noise generally decreases with increasing latitude; it also varies with time of day and season. Charts have been produced which show the estimated levels of atmospheric noise,[10] but the data on which they are based are meagre.[11] The concentration of storms in the tropics sometimes enables directional aerials to be used to reduce atmospheric noise in point-to-point systems. Background atmospherics have the general character of random noise, with a peak amplitude proportional to the square root of the receiver bandwidth; atmospherics from nearby storms, however, more closely resemble impulsive interference.

Above about 30 Mc/s the controlling factor is thermal noise in the aerial and input circuits of the receiver and valve noise in its early stages. It has a peak amplitude proportional to the square root of the receiver bandwidth, and is conveniently expressed by a noise factor, which states by how much the output signal/noise ratio is lower than that of a corresponding ideal receiver with no internal noise and an infinite input impedance, to avoid a loss of signal voltage. A receiver that matches its aerial impedance has a maximum noise factor of 3 dB, and so the aerial coupling for maximum gain is not that for minimum noise when the aerial noise is low. Typical noise factors for

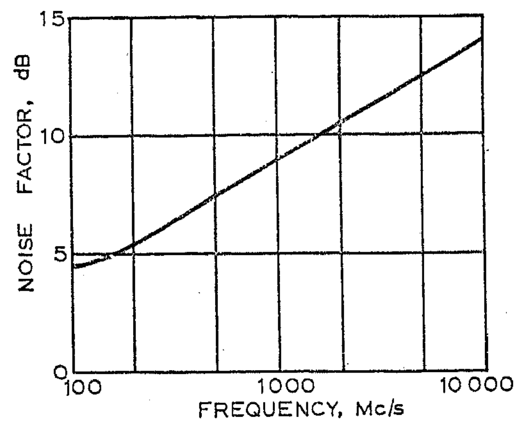


Fig. 1.—Noise factors of receivers.
From C.C.I.R. Recommendation No. 94.

receivers designed for low noise are plotted in Fig. 1; for frequencies below 100 Mc/s a figure of 4 dB can be assumed.[12]

Extraterrestrial noise has its origin in astronomical sources which include the sun, stars, interstellar gas and some invisible concentrated sources called "radio stars."[13] It is mainly significant at frequencies between about 20 and 100 Mc/s, and its magnitude may be expressed in terms of an equivalent thermal noise voltage by allotting an effective "noise temperature" to the aerial.

Interference from electrical machinery cannot in practice be completely suppressed.[14] The resulting impulsive noise differs from random noise, for the components of its spectrum are coherent; its peak amplitude is directly proportional to the receiver bandwidth, and not to its square root. The receiving stations of long-distance point-to-point radio services and of wide-band radio-relay systems are usually sited in rural areas remote from concentrations of electrical equipment. Moreover, they have directional aerials which restrict the areas from which noise is collected. They may, however, be affected by the ignition systems of motor vehicles. Impulsive noise decreases with increasing frequency, although not always so rapidly as if its voltage were inversely proportional to frequency.

(3.1.2) The Effects of Noise.

The performance of a radio system depends upon the signal/noise ratio at the input to the receiver, and the minimum acceptable ratio depends greatly upon the type of service, e.g. telegraph, telephone, broadcast relay, and upon the multiplexing and modulation techniques. Eqn. (4) shows how a bandwidth-noise exchange can be exploited by suitable methods of coding or modulation. In most modulation systems there is a critical input signal/noise ratio (the improvement threshold) below which the output signal/noise ratio falls rapidly. Thus, f.m. systems can give a much reduced output noise level compared with a.m. systems for a given input, but only for signals that exceed the peak noise level. When the noise exceeds the signal it controls the phase of the combined input, and the sidebands of the f.m. signal no longer act together to produce a large amount of frequency modulation. Again, with double-sideband amplitude modulation, when the noise level is high enough for the carrier to be fully modulated for most of the time by noise,

the output signal level is reduced.[15] Similar considerations apply to systems using pulse modulation. Single-sideband a.m. systems have no improvement thresholds, because the strong local carrier supplied to the detector provides a standard of coherence for demodulating the sidebands.

The existence of an improvement threshold affects the design of receivers. Thus, it is sometimes advantageous to use a pre-detector bandwidth much wider than is needed to accommodate the wanted signal, e.g. to allow for frequency drifts and for noise-limiting in a.m. receivers.[16] However, when the input signal is weak an increase in pre-detector bandwidth may bring the signal/noise ratio into the threshold range and cause a rise in output noise that cannot be eliminated by post-detector filtering.

Comparisons of the signal/noise ratios required for narrow-band radio systems using frequencies below 30 Mc/s have been

Table 1

System	Output circuit		R.F. input	
	Bandwidth	Signal/noise ratio	Bandwidth	Signal/noise ratio*
	kc/s	dB	kc/s	dB
*A*1 *Telegraphy*				
8 baud, low grade	1·5	−4	3	−7
50 baud, printer ..	0·25	16	0·25	2
*F*1 *Telegraphy*				
50 baud, printer ..	0·10	10	1·5	−2
*A*3 *Telephony*				
Double sideband				
(*a*) Just usable ..	3	6	6	18
(*b*) Good commercial	3	33	6	35†
Single sideband, 1 channel	3	33	3	26†

* R.M.S. signal (peak output of transmitter) to r.m.s. noise in 6 kc/s bandwidth.
† Assuming 10 dB improvement due to use of noise reducers.

published,[10,17] and some typical values are shown in Table 1 for stable conditions of propagation.

Similar comparisons for wide-band radio-relay systems are complicated by the bewildering variety of alternative methods of multiplexing and modulation, and by the large range of possible variables, e.g. number of channels, available bandwidth and length of system. Some particular comparisons have been made, but it is difficult to draw general conclusions.[18,19] Table 2

Table 2

No.	System	Bandwidth	Transmitted power
		Mc/s	watts
1	f.d.m.–f.m.	450	0·03
2	f.d.m.–f.m.	25	30
3	p.p.m.–a.m.	2 300	0·6
4	p.p.m.–a.m.	40	850

compares four hypothetical systems each designed to provide 1 000 telephone channels, 4 kc/s wide, of 60 dB signal/noise ratio, over 4 000 miles in 133 relay sections.[18] The first two systems use frequency division multiplex (f.d.m.) with frequency modulation of the radio carrier, but the first is planned for minimum transmitter power, irrespective of bandwidth, and the second for minimum bandwidth, irrespective of power. The third and fourth examples similarly compare time-division-multiplex systems using pulse position modulation (p.p.m.) with amplitude modulation of the radio carrier.

(3.2) Distortion

Distortion is the change in signal waveform that occurs within a transmission system. It is convenient to set aside changes caused by noise or extraneous signals, and to divide the remaining changes into

(*a*) Linear distortion, caused by changes in the relative magnitudes and phases of the signal components.
(*b*) Non-linear distortion, caused by circuit elements whose characteristics change with signal amplitude.

For linear distortions each signal component can be treated separately, for the superposition principle applies, but for non-linear distortions the components interact.

(3.2.1) Attenuation Distortion.

Linear distortion due to variations in the transmission loss of the system within the frequency band occupied by the signal is termed "attenuation distortion," and commonly arises from limiting the bandwidth of a radio system, either for economic reasons or to restrict interference. For telegraph, television and pulse-modulation systems the transient response of the system is of interest, and it is convenient to consider the transient response of an idealized system that transmits equally and without phase error all signal components up to a cut-off frequency, and suppresses all those of higher frequency. A signal of step waveform is distorted by such a system in two ways: its built-up time becomes finite, and ripples precede and follow the transition.

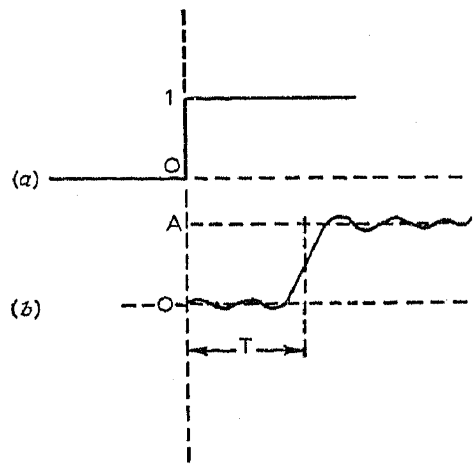


Fig. 2.—Transient response of ideal low-pass system.

(*a*) Input signal to system having uniform gain A over the frequency band 0–f c/s and zero gain beyond f.
(*b*) Output signal $= \frac{A}{\pi}\mathrm{Si}\,[2\pi f(t-T)]$ where T = transmission time.

The output waveform illustrated in Fig. 2 is described by the sine integral function[9,20]

$$\mathrm{Si}\,(x) = \int_0^x \frac{\sin x}{x} dx \quad . \quad . \quad . \quad . \quad . \quad (5)$$

and for a system having a uniform gain, A, over the frequency band 0–f c/s, and a uniform transmission time T sec, the output signal produced by a unit step at the input is

$$\frac{A}{\pi}\mathrm{Si}\,[2\pi f(t-T)] \quad . \quad . \quad . \quad . \quad . \quad (6)$$

The output signal builds up from 10 to 90% of its final amplitude in approximately $0 \cdot 45(1/f)$ sec. The ripples have maxima separated by approximately $1/f$ sec, and the maximum of the first ripple preceding or following the transition corresponds to an overshoot of approximately 9%. The response to rectangular pulses can be determined by adding the separate responses to their edges. Fig. 3 indicates the response of an idealized system

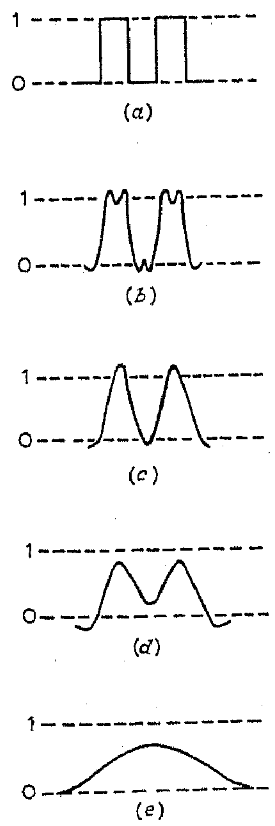


Fig. 3.—Effect of system bandwidth on pulse response.

(a) Input signal.
(b) Output signal $f = 2(1/T)$.
(c) Output signal $f = 1/T$.
(d) Output signal $f = \frac{1}{2}(1/T)$.
(e) Output signal $f = \frac{1}{4}(1/T)$.

to a symmetrical pair of pulses. The minimum bandwidth for the resolution of adjacent pulses, e.g. telegraph elements or details in a television picture, corresponds approximately to Fig. 3(d), i.e. $f = 1/(2T)$, where T is the pulse duration. The transient response of a band-pass system can be derived, using the superposition principle, by subtracting the responses of two low-pass systems with cut-off frequencies corresponding to the upper and lower limits of the pass band, as illustrated in Fig. 4. For double-sideband a.m. signals, Figs. 2, 3 and 4 apply when

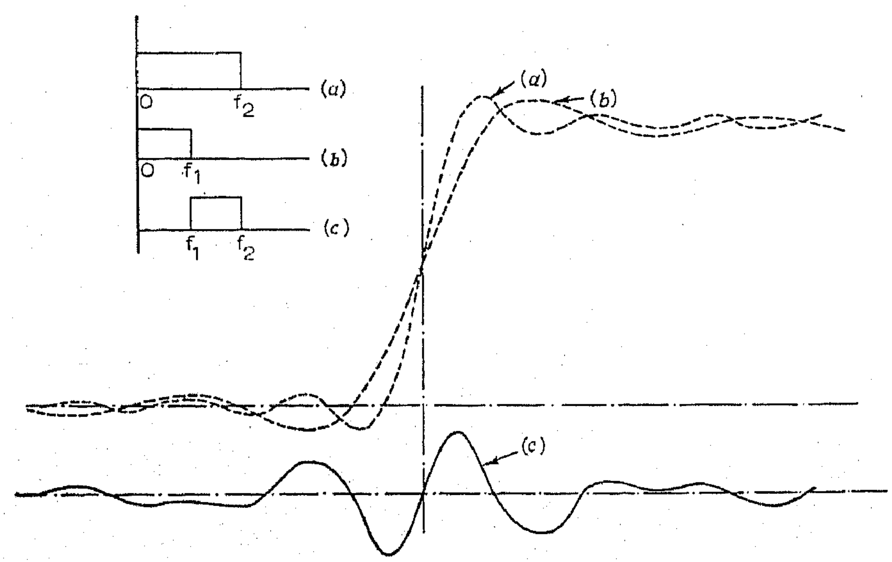


Fig. 4.—Superposition principle and transient response.

the pass band extends equally by $\pm f c/s$ about the carrier frequency. When the carrier frequency does not coincide with the centre of the pass band the upper and lower sidebands become asymmetrical. This introduces an additional type of distortion, known as quadrature distortion, which depends upon the degree and nature of the sideband asymmetry; general conclusions cannot be presented, although particular analyses have been published.[9,21] The filling-in of the lower corners and the peaking of the upper corners of square-wave signals are typical effects of quadrature distortion.

Frequency modulation differs from amplitude modulation in that the bandwidth occupied by the signal depends upon the depth of modulation. The distortion produced by limiting the bandwidth of an f.m. system is therefore worse at the higher modulating signal amplitudes, and non-linear distortion is introduced. The effect of bandwidth limitation depends upon its ratio to the frequency deviation,[22,23]; Fig. 5 illustrates

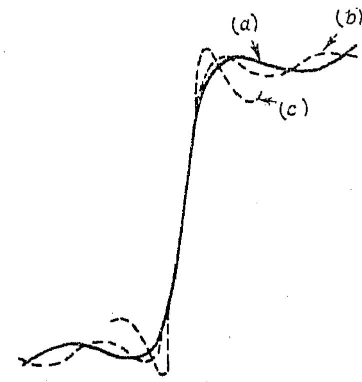


Fig. 5.—Transient response of f.m. systems of restricted bandwidth.

(a) A.M. system for comparison: equal bandwidth.
(b) F.M. system: deviation = $\frac{1}{8}$ × bandwidth.
(c) F.M. system: deviation = $\frac{1}{3}$ × bandwidth.

the step-wave responses of two f.m. systems, with an a.m. response for comparison. The initial rate of rise is the same for all systems, and when the bandwidth is four or more times the peak deviation it depends only upon the filter bandwidth. When the f.m. carrier is not located at the centre of the pass band the transient fluctuations are increased.

Attenuation distortion can be assessed by the paired-echo analysis,[9,24] and in a system having no phase error and an amplitude characteristic represented by

$$A(\omega) = a_0/2 + a_1 \cos c\omega \quad . \quad . \quad . \quad . \quad (7)$$

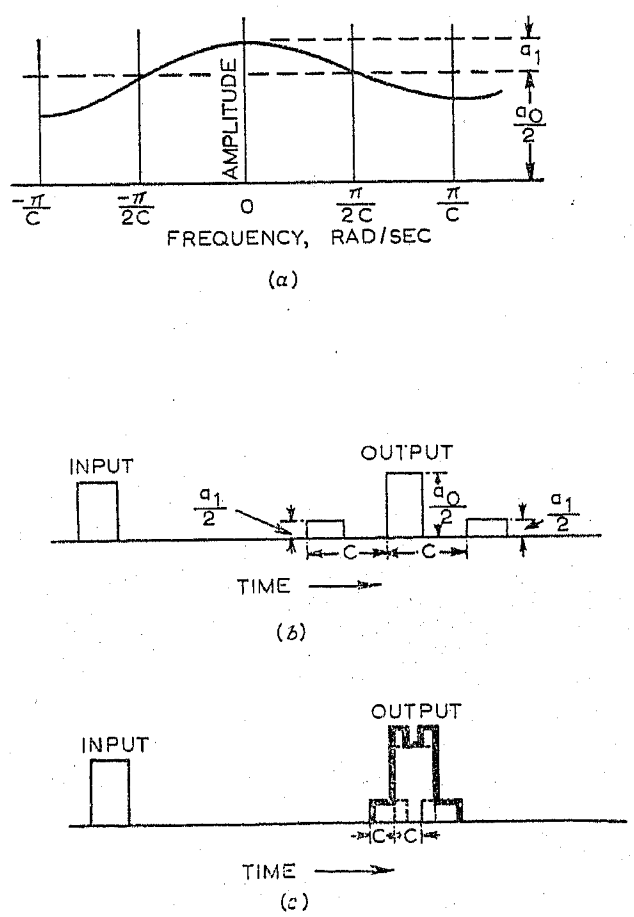


Fig. 6.—Wheeler "paired echo" analysis of attenuation distortion.

(a) System amplitude/frequency characteristic.
(b) Widely spaced echoes.
(c) Closely spaced echoes.

(where a_1 is small compared with a_0) the output signal consists of the main undistorted signal plus two weaker echoes similar in shape to the undistorted signal, one preceding the signal and the other following it at equal intervals of c sec. The echoes each have an amplitude of a_1/a_0 times that of the main output signal,

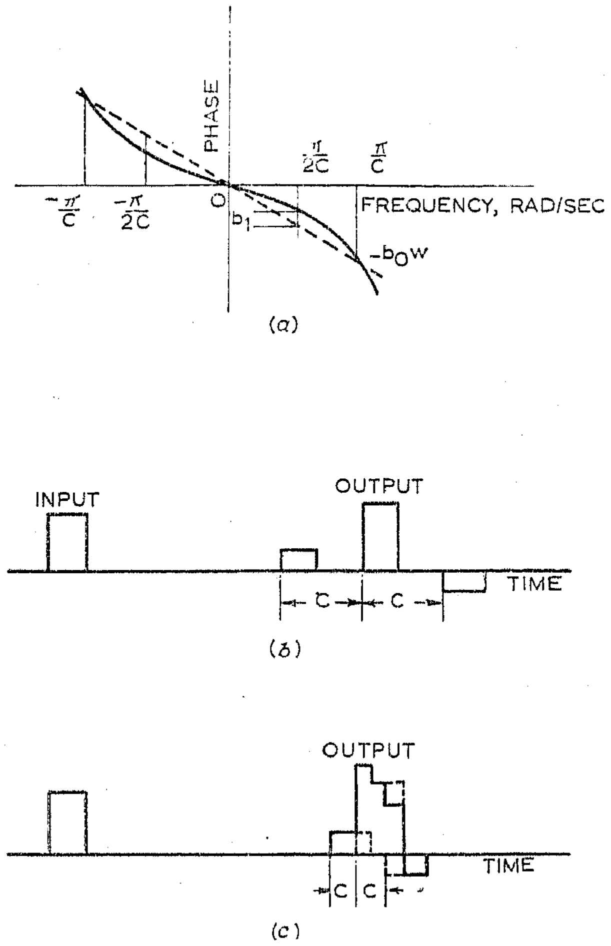


Fig. 7.—Wheeler "paired echo" analysis of phase distortion.

(a) System phase/frequency characteristic.
(b) Widely spaced echoes.
(c) Closely spaced echoes.

as illustrated in Fig. 7. Any attenuation/frequency characteristic can be synthesized from a series of cosinusoidal harmonic components each with its appropriate pair of echoes.

(3.2.2) **Phase Distortion.**

Distortionless transmission requires the components of a complex signal to arrive simultaneously, i.e. with their relative phases unaltered. In the transmission of a recurrent waveform of any shape it would make no difference if the phase shift, B rad, at the frequency, ω rads/sec, of any component were increased by an integral multiple of 2π rad, since successive cycles of each component of the signal are indistinguishable: i.e. it is sufficient if over the whole frequency range occupied by signal components

$$B = \omega T + 2n\pi \text{ radians} \quad . \quad . \quad . \quad . \quad (8)$$

where T is the transmission time. However, when the phase characteristic is represented by

$$B = \omega T + 2n\pi + \phi \text{ radians} \quad . \quad . \quad . \quad . \quad (9)$$

i.e. by a straight line intercepting the zero-frequency axis at an ordinate $(2n\pi + \phi)$, the first two terms correspond to distortionless transmission and may be subtracted, leaving a constant phase shift of ϕ rad at all frequencies, which, by displacing the signal components, causes considerable waveform distortion. For a.m. waves the zero-frequency phase intercept is unimportant; a linear phase/frequency characteristic of the form given by eqn. (9) is sufficient.

Two constant quantities having the dimension of time can be defined, namely

(a) B/ω, the average slope of the phase-frequency characteristic, called the phase delay.
(b) $dB/d\omega$, the differential slope of the phase-frequency characteristic, called the differential delay.

A non-linear phase/frequency characteristic shows variations in both quantities and either may be used for specification. Differential delay is the more suitable for modulated waves, the distortionless requirement being that the differential delay should be constant. In the absence of appreciable attenuation distortion the differential delay at any frequency is the time delay experienced by the maximum value of the envelope of a signal element corresponding to a small group of signal components having frequencies close to the frequency in question. For this reason differential delay is sometimes called group delay or envelope delay. When there is appreciable attenuation distortion the results are harder to interpret.[25]

Phase distortion also can be assessed by the paired-echo method, for in a system with no attenuation distortion and a phase/frequency characteristic of the form

$$B = -b_0\omega + b_1 \sin c\omega \quad . \quad . \quad . \quad . \quad (10)$$

(where b_1 is small compared with b_0) the output signal consists of the main undistorted signal plus two weaker echoes similar in shape to the undistorted signal, preceding and following the signal at equal intervals of c sec. The two echoes each have amplitudes of $[J_1(b_1)/J_0(b_1)]$ times the undistorted signal, where $J_0(b_1)$ and $J_1(b_1)$ are Bessel functions of the first kind, of zero and first order respectively, and the following echo is a negative one, as illustrated in Fig. 7. It has been estimated[9] that a pair of adjacent pulses ceases to be resolved when the variation from a linear phase/frequency characteristic approaches one radian in a frequency range of $1/(3T)$ c/s, where T is the pulse duration in seconds. This is only a rough guide, for a large phase deviation inside a narrow frequency range causes little distortion, unless that range happens to contain a large proportion of the signal energy. Like attenuation distortion, the overall effect of phase distortion is non-linear in f.m. modulation systems.[26,27]

(3.2.3) **Non-Linear Distortion.**

Linear distortions can be corrected by passive attenuation or phase equalizers. Non-linear distortions, however, generate additional signal components and cannot be easily corrected.

Non-linear distortion arises in a radio system when the amplitude of the output signal is not directly proportional to that of the input signal; usually the relation between the two amplitudes can be represented by a power series, and simple trigonometry shows that for an input signal consisting of a single component the output signal is a harmonic series. When the signal has several components of frequencies ω_1, ω_2, ω_3, etc., then as well as harmonics of each of these frequencies the output contains intermodulation components having the frequencies $(m\omega_1 \pm n\omega_2 \pm p\omega_3 \pm \ldots)$ where $m, n, p \ldots$ are integers. Similar results are obtained for a system with an input/output relationship in the form of a power law,[28] e.g. $r = kV^{\gamma}$.

The effects of harmonic and intermodulation components can sometimes be assessed by treating them as noise, and methods have been described for calculating this intermodulation noise in both a.m. and f.m. systems.[29–32]

In a.m. radio systems, and in parts of other systems carrying the unmodulated signals, non-linear distortion is due mainly to curvature of the dynamic characteristics of thermionic valves in amplifiers and modulators. The distortion can be controlled

either by suiting the operating conditions to the valves used or by negative feedback.[33]

In f.m. and p.m. radio systems non-linearity mainly arises from attenuation and phase distortions, and these can be reduced by equalizing networks.[34] The relationship between echo signals and attenuation and phase distortions has been mentioned, and echoes caused by multi-path propagation or by reflections in aerial feeders are important sources of non-linear distortion in f.m. systems.[32]

Harmonic generation changes the signal waveform, but relatively large changes can often be tolerated. Telegraph systems are usually intentionally non-linear, for they use amplitude limiters. Telephone signals can also be limited in amplitude, or clipped, to the point where virtually nothing of the speech wave remains but an indication of the instants at which it crosses the zero axis, and although this distortion is audible it does relatively little to reduce intelligibility.[35] Non-linear distortion in television and facsimile systems disturbs the tone range of the received pictures and in television alters the picture/synchronizing-signal ratio; an appreciable amount of non-linear distortion can be tolerated, but little has been published to indicate exact limits.[36,37] Harmonics generated in radio transmitters may cause widespread interference with other radio services, particularly those generated by the high-powered transmitters used for long-distance point-to-point services below 30 Mc/s. The maximum amount of harmonic power supplied to the aerial, however, is strictly limited by international regulation.

Mutual interference between two or more signals simultaneously present in non-linear parts of a radio system cannot subsequently be eliminated. It is particularly important in multi-channel systems, where it produces crosstalk between the channels and is often the limiting factor.[29,31] Non-linearity in the early stages of a radio receiver may cause the wanted and any unwanted signals to intermodulate before the unwanted signals have been suppressed by the tuned circuits of the receiver. This commonly transfers the modulation of the unwanted signals to the wanted signal, an effect known as cross-modulation. Moreover, strong signals may generate intermodulation products with carrier frequencies represented by $(m\omega_1 \pm n\omega_2 \pm p\omega_3 \pm \ldots)$ when the first stages of the receiver are inselective, or a wide-band aerial amplifier is used, or non-linear resistances exist in aerial or earth systems or in nearby metalwork.[38,39] Intermodulation occurs at transmitting stations when coupling between the aerials of different services causes several signals to be present simultaneously in the non-linear final stages of the transmitters.[40] For systems using frequencies below 30 Mc/s, non-linearity can also occur in the ionosphere in the immediate neighbourhood of a powerful low-frequency transmitter, as in the well-known Luxembourg effect.[41]

In the relaying of American colour-television signals it is important that the phase-shift experienced by signals near the colour sub-carrier frequency should be substantially independent of signal amplitude;[42] this is a new type of linearity requirement for radio systems.

(3.3) Radio-Wave Propagation

(3.3.1) Signal Strength.

The propagation of radio waves has been the subject of much theoretical and practical investigation. Alternative modes of propagation exist: by direct waves, and by waves reflected, refracted, diffracted or scattered from the ground and in the troposphere and ionosphere, and each has an importance depending upon frequency. Frequency is a major factor, with a broad division at 30 Mc/s between lower frequencies, suitable for long-distance propagation, and higher frequencies, limited to ranges of some 30–300 miles. The influences of solar, seasonal and meteorological conditions are known in outline.[10,43,46]

What a power engineer would consider to be an entirely negligible fraction of the power radiated by the transmitter, reaches and influences the remote receiver. Thus, the measured insertion loss between the aerials of a 13 Mc/s Britain–America radio circuit varied between 120 and 180 dB, so that at best only 1 part in 10^{12} of the transmitted power was usefully employed.

Radio systems ideally require strong and constant signals, without the need for excessive transmitter power. The approximate strength of the signals set up at a distant point can be estimated, but the errors increase when alternative modes of propagation are possible. Constancy of signal strength is beyond control, and variations have to be countered by equipment design.

(3.3.2) Fading.

Fading is a characteristic of received radio signals of all frequencies, and may be non-selective (when the amplitude of the signal as a whole varies, usually quite slowly) or selective (when components of the signal having adjacent frequencies fade independently at a relatively rapid rate). Selective fading occurs when the signal is simultaneously received over two or more paths which are changing in length, and as a variable echo it gives rise to changing attenuation and phase distortions, which produce severe non-linear distortion in some types of radio system.[47] Thus, in double-sideband a.m. systems, when the carrier fades selectively it produces the effect of severe over-modulation. Frequency-shift telegraph signals received simultaneously over two paths exhibit a severe "spiky" distortion. Continuous-wave telegraph signals have their durations increased, and at the fading minimum "dumb-bell" distortion may split the pulses.

When the minimum signal/noise ratio is adequate, non-selective fading can be countered by automatic gain control (a.g.c.) circuits in the radio receiver. In f.m. and telegraph systems amplitude limiters can supplement the a.g.c. circuits.

Selective fading presents a more difficult problem, and various methods have been devised to minimize its effects, e.g. directive aerials (which select signals from only one of the possible propagation paths), diversity methods (which provide two or more signals fading independently so that at least one strong signal is always present) and specially tolerant signalling systems. The multiple unit steerable antenna (m.u.s.a.)[48] is an example of an anti-fading directive aerial.

Diversity methods[49,50] include

(*a*) Frequency diversity, i.e. sending the same signal on a number of frequencies, including some obsolete techniques, e.g. damped and interrupted c.w. emissions, which spread the signal over a wider frequency band than the minimum necessary for the speed of signalling; their modern counterpart is "phase scintillation" in frequency-shift telegraph systems.

(*b*) Spaced-aerial diversity, in which two or three separate aerials are spaced one or more wavelengths apart, either parallel or perpendicular to the propagation path, and the strongest received signal is automatically selected.

(*c*) Polarization diversity, which resembles (*b*) but uses coincident aerials responding to waves having mutually perpendicular planes of polarization.

(*d*) Arrival-angle diversity, e.g. separate receivers attached to a m.u.s.a. system set to receive separate downcoming rays, and the strongest signal selected. Azimuthal angle-diversity has been used for receiving signals scattered from the ionosphere.

(*e*) Time diversity, in which the message is repeated after a short interval and the two received messages compared, i.e. telegraph Verdan working.

Selective fading may be offset by system design. Thus, in a frequency-shift telegraph system selective fading produces severe beat-frequency disturbances near the start and finish of each signal element. However, the consequences can be minimized

by reducing the signalling speed until these disturbances occupy only a negligible proportion of each signal element. Again, single-sideband reduced-carrier systems provide a much less distorted signal in the presence of selective fading than do double-sideband a.m. systems, and are widely used for long-distance radiotelephony.

The allowance to be made for fading has been studied by the International Consultative Radio Committee (C.C.I.R.) and provisional allowances for different 3–30 Mc/s radio systems have been published.[51] These not only take account of fading but also allow uncorrelated day-to-day fluctuations of 10 dB in the intensities of both signal and atmospheric noise. The conclusions so far reached by the C.C.I.R.[52] are provisional, but for signals between 3 and 30 Mc/s they point to a Rayleigh distribution of amplitudes for a fading single-frequency signal observed over periods of a few minutes, and to a log-normal distribution over periods of 15 min to one hour, as shown in Fig. 8.

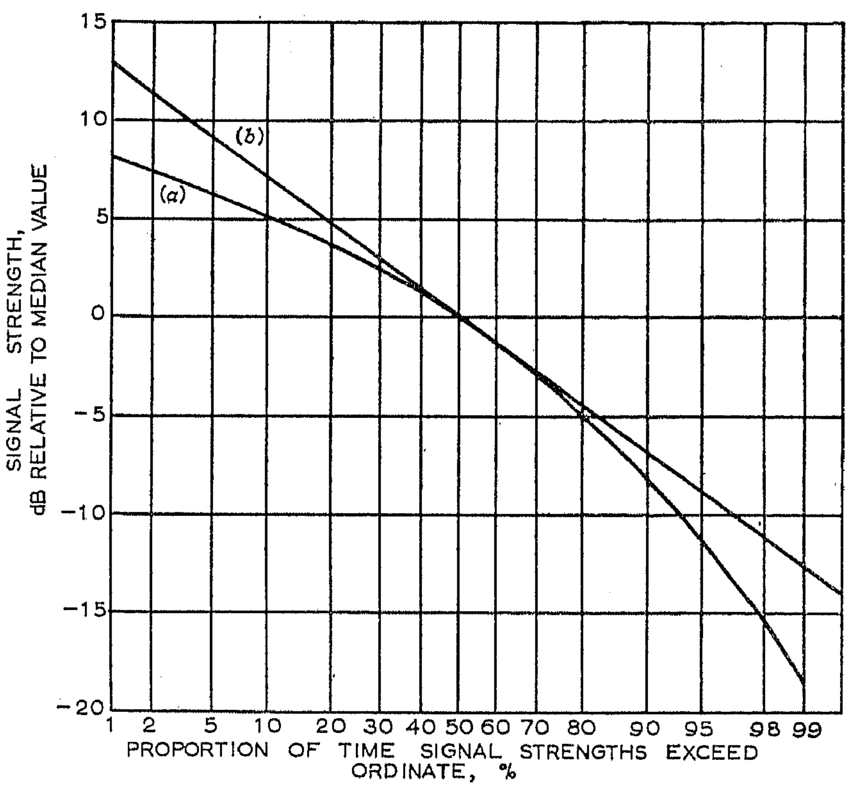


Fig. 8.—Amplitude distributions for fading single-frequency signals in the 3–30 Mc/s range.

(*a*) Rayleigh distribution (periods of a few minutes).
(*b*) Log-normal distribution (periods of 15–60 min.)

Fading of signals with frequencies above 30 Mc/s is largely due to tropospheric reflections and changes in ground clearance caused by variations in the refractive index of the atmosphere. Some results[53–55] indicate a Rayleigh distribution of signal amplitudes. Signals between about 30 and 100 Mc/s are regularly propagated over distances of the order of 1 000 miles by scattered reflections from ionospheric irregularities, although the signals are weak and fluctuate rapidly;[56] signals on frequencies above 30 Mc/s can also be received at ranges of 100–300 miles by scattering from irregularities in the troposphere.[57]

(3.4) Interference

The frequency bands allocated to point-to-point radio systems below 30 Mc/s are to-day extremely congested, and mutual interference arises because adjacent channels cannot be separated as widely as theoretical considerations require. Ideally, the factors listed below would all receive due weight, but in practice the minimum frequency-separation is usually determined by trial.

(*a*) Required signal/interference ratios.
(*b*) Necessary bandwidths of the adjoining systems.
(*c*) Transmitter frequency stability.
(*d*) Out-of-channel radiations.
(*e*) Receiver bandwidth necessary.
(*f*) Rate of increase of attenuation outside receiver pass-band.
(*g*) Receiver frequency stability.
(*h*) Aerial directivity at transmitter and receiver.
(*i*) Allowances for fading and slower changes in signal strength.

Similar problems arise above 30 Mc/s in radio-relay systems, and here it has been shown[18] that it is not necessarily those systems having the narrowest bandwidth that can accommodate the greatest number of channels in a given band. Some wider-band systems can accept much more interference from similar systems than the narrower-band systems, but this consideration applies only when all the mutually-interfering systems are of similar type.

The required signal/interference ratios depend upon the type of service, e.g. telegraphy, telephony, etc., and the method of modulation, but although this question is being studied by the C.C.I.R. no recommendations have yet been made and much work remains to be done.

Frequency instability at the transmitter and the receiver means that the signals can cause and suffer interference over a wider frequency band than the minimum necessary for the type of service.[58] The increase in effective bandwidth is proportionately greater with narrow-band signals, and this suggests one advantage of wide-band multiplex emissions.[59,60] The frequency tolerances prescribed internationally for point-to-point radio services jump from 0·003 to 0·02% at 30 Mc/s and rise to 0·75% above 500 Mc/s, which illustrates the fact that interference between radio services is at present mainly a problem for the relatively narrow range of frequencies capable of long-distance propagation.

Two point-to-point radio services using frequencies below 30 Mc/s may be able to use the same frequency simultaneously by the use of directive aerial systems, which also serve to reduce interference from services on adjacent frequencies and to decrease the transmitter power necessary to provide the necessary margin over receiver noise. The aerial directivity achieved in practice has been limited by economic considerations, for existing highly directive aerials are several wavelengths in size. Theory indicates that large size is not essential, and indeed the lack of correlation between the signals induced in aerials a wavelength or more apart suggests that it may be a handicap. It is difficult to design a directive aerial that is adequately free from undesired responses in directions other than the main beam. Again, radio signals propagated by reflection from the ionosphere sometimes arrive from directions that differ appreciably from the great-circle path to the transmitter, and a sharply directional aerial might attenuate such signals and degrade the circuit.

Receiver selectivity[61] is clearly important in reducing interference from services operating on adjacent frequencies, but a high degree of total selectivity is valueless unless there is sufficient selectivity in the early stages of the receiver to prevent cross-modulation or blocking. Unwanted responses may also be produced by non-linearity in the early stages, so giving rise to intermodulation between two strong signals and producing sum or difference frequency products falling in the receiver pass-band or in the i.f. pass-band of a superheterodyne receiver. In superheterodyne receivers the responses to a variety of unwanted signals also must be reduced to acceptable values; these include image-frequency and intermediate-frequency signals, and in poor receivers the responses to signals having frequencies given by

$$f = \frac{n}{m}f_s + \frac{n \pm 1}{m}f_i \quad . \quad . \quad . \quad . \quad . \quad (11)$$

when the first beating oscillator frequency is $(f_s + f_i)$, where f_s and f_i are, respectively, the wanted signal and intermediate frequencies.

(4) GENERAL PROCESSES

Some processes are common to many kinds of point-to-point radio system. At the sending end the messages are converted into electrical signals of appropriate form by a process that can be regarded as coding. In many systems a single set of radio equipment carries many signals simultaneously, the several electrical signals being combined for transmission by a variety of multiplexing methods. After multiplexing, the combined signals are modulated onto the radio frequency carrier wave. At the receiver the converse processes of demodulation, demultiplexing and decoding take place, as well as other processes intended to offset the disturbances introduced during the propagation of the signals from the transmitter to the receiver.

(4.1) Coding

The development of radio-teleprinter systems has led to the international standardization of the 5-unit code, using international telegraph alphabet No. 2. Five-unit code systems have the disadvantage that any corruption of the signal is liable to produce a false character, since all possible combinations of elements are used. When redundancy is introduced by using codes containing more than five units it can be arranged that a single disturbance of a character, e.g. the changing of one mark element to a space, does not simulate another character but produces an obvious error. In more complex codes a single error leaves the mutilated character more like the original character than any other, and the error can be detected and corrected at the receiver; still more complex codes will accept two or more errors in a single character.[7,62] A practical limit is set to the number of units in telegraph codes that have been used for point-to-point radiocommunication by the reduction of working speed in a given bandwidth as the number of code elements is increased.[63]

Redundancy can be introduced by repeating the message after a short interval; this method, known as Verdan working, is equivalent to the use of a code having twice as many units. Again, the use of more than two signalling conditions can give redundancy; thus a 5-unit 3-condition code has nearly as many $[(3^5 - 32) = 211]$ redundant symbols as an 8-unit 2-condition code $[(2^8 - 32) = 224]$. However, multi-condition codes require higher signal/noise ratios than 2-condition codes for a given liability to error.

Error correction can be obtained without a lengthy code when a return radio channel is available, by using a 7-unit error-detecting code. The method is to return a special signal to the sending end requesting the repetition of the last few characters whenever an error is detected, and equipment of this type is coming into use on point-to-point radio circuits. The use of different telegraph codes on interconnected line and radio systems requires devices converting automatically from one code to another.

In point-to-point radiotelephone systems the signals from the land-line may be subjected to various "coding" processes before application to the radio transmitter. First, the average levels of the speech signals from different subscribers differ widely at the radio terminal,[64] and a major saving in transmitter power can be made if a high and fairly constant depth of modulation is ensured by using an automatic variable-gain amplifier to maintain a constant speech-signal level. Again, many single-sideband-receiver a.g.c. circuits respond to variation in the level of the reduced carrier and have time-constants of several seconds to carry over selective fades of this carrier. Speech signals from such receivers are subject to rapid level variations of up to 20 dB, which may be corrected by a second constant-volume amplifier.

The effects of radio noise, interference and crosstalk between the channels of a multi-channel system can be reduced by expanding the volume range of the received speech with an amplifier of automatically adjusted gain, which amplifies strong signals more than weak ones. Volume expansion distorts the speech waveform, but this distortion can be avoided by an inverse process of compression at the sending end; the combination of compressor and expander is known as a "compandor."[65] These compandors respond to level changes occurring in periods of the order of 15–20 millisec, and are accordingly called syllabic compandors. Signals compressed by a syllabic compandor can be sent in the bandwidth occupied by the original signals without significantly increasing the attenuation and phase requirements, and syllabic compandors can therefore be added to existing systems. In another class of compandor, volume compression occurs sensibly instantaneously[66] and produces harmonic and intermodulation products extending over several additional octaves, all of which must be transmitted to the expander for the original speech to be recovered without distortion. Since the compandor does not alter the information content of the signal it is theoretically possible to recode the compressed signals so that they can be sent without loss and without increase of bandwidth,[67] but this would generally involve severe attenuation and phase requirements. Instantaneous compandors could conveniently be used in pulse-modulation systems, for no extra bandwidth is required to transmit the compressed *samples* of speech and the usual companding advantages would be obtained.

In a radiotelephone system the average depth of modulation is limited by the high ratio of instantaneous peak to r.m.s. voltage—12–15 dB for normal speech. The average depth of modulation can be increased almost to 100% by severely clipping the speech signal with amplitude limiters, a process that leads to very little loss of intelligibility.[35] Another method of saving transmitter power in radiotelephone systems depends on attenuating the lower-frequency components of the speech signal before they are applied to the transmitter—a process called pre-emphasis, an inverse, de-emphasizing, network being used at the receiver output. This method is particularly useful in f.m. systems, with their well-known triangular noise spectrum. It is sometimes adopted in multi-channel f.m. point-to-point radio systems using frequency-division multiplexing to improve the signal/noise ratios in the higher-frequency channels.

Some speech coding methods reduce bandwidth by reducing the redundancy of ordinary speech signals when intelligibility is the criterion rather than fidelity. In the vocoder[68] system the voice sounds are analysed and a group of relatively-slowly-varying signals are transmitted in a bandwidth of some 400 c/s to control a synthetic speech reproducer. In another method[69] the speech frequency-range is compressed before transmission by recording it and scanning the record with moving pick-ups. An inverse frequency-expansion process is used at the receiver, and reductions in the r.f. bandwidth by two or four times are practicable. Both methods affect the naturalness of the reproduced speech, and neither has yet been used commercially.

The clearest example of coded radiotelephone signals is found in pulse code modulation (p.c.m.), which represents the speech signals by a series of on-off pulses of fixed amplitude. The receiver is required only to recognize whether a pulse is present or not, which can be done even when the accompanying noise and interference amount to about 50% of the pulse peak amplitude. Again, the on-off pulses can be cleaned and reshaped by regenerator circuits to prevent the accumulation of distortion along a point-to-point radio-relay system. The cost of these advantages is complexity and large bandwidth.[18]

The scanning processes used for the transmission of still and moving pictures are examples of coding in the broad sense of converting the message into a form suitable for transmission

over a radio channel. Scanning produces a periodic signal with its energy concentrated near harmonics of the scanning frequencies,[70] and there are limited possibilities of accommodating other signals in the gaps in the spectrum between the harmonics—a fact that is exploited in the American system of colour television.[71] It is also possible in principle to adapt the scanning processes to take account of the large amount of redundancy in the average picture,[72–74] but no advantage has so far been taken of this possibility.

(4.2) Modulation and Demodulation

The choice between systems of modulation depends upon the type of message signal, the transmission obstacles to be surmounted and economic factors.

Amplitude modulation systems can be divided according to whether the sideband distribution is symmetrical or asymmetrical about the carrier wave. The most important asymmetrical-sideband systems are the independent-sideband systems used for long-range radiotelephony between 3 and 30 Mc/s. In independent-sideband systems it is usual to reduce the carrier wave considerably and to supply a large local carrier to the demodulator, which reduces the distortions produced by the suppression of one sideband and by selective fading, and improves the performance at low signal/noise ratios. A total gain in signal/noise ratio of 9 dB is possible in comparison with a double-sideband system of equal peak power.[75] Single-sideband signals have been generated either by removing the carrier and the unwanted sideband by filter circuits, usually at an intermediate frequency,[76] or by out-phasing.[77–79] The out-phasing technique is the easier when it is necessary to transmit very-low-frequency signals, e.g. for music circuits.

An unwanted signal on an adjacent frequency gives rise to interference in an a.m. system from the heterodyne note between the two carrier waves, and from the demodulation of the unwanted signal against its own carrier and against the carrier of the wanted signal. When a linear envelope-detector is used and one of the two signals is stronger than the other the modulation of the weaker signal is apparently suppressed.[9] This modulation suppression is important in reducing intelligible interference, although it does not affect the inter-carrier beat or the unintelligible crosstalk produced by demodulation against the wanted carrier.

Single-sideband systems do not exhibit the threshold effect[80] mentioned in Section 3.1.2, by virtue of the strong local carrier used at the detector, and a similar technique of coherent detection is sometimes used for very weak double-sideband a.m. signals.[81] Correlation methods could also be used when the signals are periodic.[82]

Frequency modulation and phase modulation are closely related, and are often combined in the same system. When the modulating signal consists of a single frequency, the only difference between frequency-modulated and phase-modulated signals is in the degree of modulation. For multi-component signals the two systems differ, because the phase-modulated signals have much larger frequency deviations at high modulating frequencies. In frequency or phase modulation a single modulating frequency produces an infinite series of pairs of sidebands spaced from the carrier at multiples of the modulating frequency, and when several modulating frequencies are present pairs of sidebands are also produced corresponding to all the sum and difference frequencies of the form $(mf_1 \pm nf_2 \pm pf_3 \pm \ldots)$ where $m, n, p, \ldots$ are integers. The loss of any of these sideband components by the use of too restricted a bandwidth will produce non-linear distortion and intermodulation between the modulation components. Noise and interference have a random phase relationship to the f.m. carrier, and when the frequency deviation considerably exceeds the highest modulating frequency most of the noise and interference can be eliminated by low-pass filters after the f.m. detector.[83–85]

The principal applications of frequency and phase modulation to point-to-point radio systems are

(*a*) Frequency-shift telegraph systems (3–30 Mc/s).
(*b*) Short-range telephone links (30–300 Mc/s).
(*c*) Wideband radio-relay systems for multi-channel telephony or television (100–10 000 Mc/s).

Since frequency-shift telegraph systems signal only two conditions, the signal may be detected either by a conventional f.m. discriminator or by a pair of a.m. detectors tuned to the separate marking and spacing frequencies. These alternative methods behave differently in the presence of noise and selective fading, and it has not yet been clearly established which method is superior.[86–90] Short-range radiotelephone links are used in the United Kingdom for communication to islands near the mainland. Simple single-channel links of this type use phase modulation in order to simplify the frequency-control arrangements,[91] and frequency modulation is used in the more elaborate multi-channel links to obtain a high standard of performance with comparatively small transmitter powers.[92] Wide-band relay systems use frequency rather than amplitude modulation to avoid non-linear distortion due to curvature of valve characteristics.[34]

Pulsed carriers can be modulated in amplitude, frequency or phase.[93] They can also be modulated in shape, e.g. rectangular pulses can be modulated in width and trapezoidal pulses can be modulated in build-up time.[94] Pulse code modulation can also be used.[95,96] Pulse-amplitude-modulation systems cannot exchange bandwidth for improvements in signal/noise ratio like other pulse methods; their advantage is simplicity of equipment. Pulse-phase-modulation systems also use relatively simple equipment but need wide frequency bands, and have been little used so far. Pulse-code-modulation systems need wide bands, but their binary code elements are very tolerant of interfering signals and noise, and can be regenerated from time to time. Although these systems are attractive they have not so far been applied on any large scale to point-to-point radiocommunication, but they have been extensively studied.

(4.3) Multiplexing

Four methods of multiplexing have been principally used,[18,19] namely

(*a*) Frequency division.
(*b*) Time division.
(*c*) Carrier-phase division.
(*d*) Signalling-condition division.

In frequency-division multiplex (f.d.m.) the individual signals are modulated on different sub-carriers and combined and separated by wave filters. This system, when single-sideband suppressed-carrier amplitude modulation is used, occupies the smallest possible total bandwidth, and is often used as a standard of comparison. It is, however, more susceptible to noise and interference than some alternative systems which occupy wider frequency bands. The principal advantage of f.d.m. systems for radiotelephony comes from the fact that the various channel sidebands rarely peak together, so that the required peak transmitter power increases only slowly as the number of channels increases.[18] A further advantage comes from the derivation of the sub-carrier frequencies from a single, stable source, which allows adjacent channels to be packed tightly without wasting

bandwidth on "guard band" allowances for frequency instability. This is the principal advantage in radiotelegraph f.d.m. systems, for in these the non-simultaneous-load advantage does not apply. A major limitation to the use of frequency-division multiplexing with amplitude-modulation for radio services is amplitude non-linearity of the transmitter leading to interchannel crosstalk.

Time-division multiplex (t.d.m.) makes the full power of the radio transmitter available to transmit a sample of each channel in turn for equal, brief periods; it is particularly easily adapted to telegraph and pulse-modulation systems.[97] The theoretical minimum total bandwidth of a t.d.m. system is the same as that of an equivalent f.d.m. system. Transmitter non-linearity does not lead to interchannel crosstalk in t.d.m. systems, but crosstalk arises from factors that increase the duration of the individual pulses causing them to overlap, e.g. insufficient band-width.[98] T.D.M. systems involve the radio transmission of short pulses and are subject to distortion and crosstalk under conditions of multi-path propagation. Since each channel in turn uses the full transmitter power the non-simultaneous-load advantage is not obtained in t.d.m. telephone systems; however an instantaneous compandor could be used. T.D.M. systems need not be synchronous, and a system has been proposed in which each channel is sampled at irregular intervals, although at the same average rate for all channels;[99] the magnitude of each sample is conveyed by a group of pulses which are coded to identify the channel. Such a system has no rigid limit to the number of channels provided, but crosstalk increases as more channels are added. It has been suggested for rural radio-telephone exchanges.

Carrier-phase-division multiplex systems are comparatively rare; in the most familiar example two different signals separately modulate the amplitudes of two carriers of the same frequency, but in quadrature phase relationship, namely the sine and cosine carriers. The carrier waves are suppressed and the sideband signals are mixed and transmitted together. At the receiver the signals are applied to two demodulators, one supplied with a sine and the other with a cosine carrier from a local synchronized source.[8] If the attenuation and phase characteristics of the signal path between the modulator and demodulator are symmetrical about the carrier frequency the two signals can be recovered separately without crosstalk. Single-sideband transmission is impossible and the system offers no saving in bandwidth compared with single-sideband systems, although it is useful, as in television systems, where the very low frequencies of some signal components make it impossible to construct filters to select the sideband and reject the carrier.

Signalling-condition-division multiplex is exemplified by the Twinplex telegraph system, in which two telegraph signals are sent simultaneously by one transmitter by allocating four signalling conditions, to represent the four combinations of marking and spacing signals on the two channels.[100] The full transmitter power is available to each channel, although if the radio system is troubled by noise or interference the effects on the two channels will generally be different. This system can be extended to more than two channels, although the number of signalling conditions required increases rapidly as the number of channels is increased.

(5) CONCLUSION AND ACKNOWLEDGMENT

This paper is intended to serve as an introduction to the references quoted, but these, although numerous, are inevitably no more than a sample of the extensive literature of point-to-point radio systems.

The author's thanks are due to the Engineer-in-Chief of the General Post Office for permission to publish the paper.

(6) BIBLIOGRAPHY

References marked * relate to systems or equipment established in practical use.

(1) GABOR, D.: "Theory of Communication," *Journal I.E.E.*, 1946, **93**, Part III, p. 429.

(2) SHANNON, C. E., and WEAVER, W.: "The Mathematical Theory of Communication" (University of Illinois Press, 1949).

(3) CHERRY, E. C.: "A History of the Theory of Information," *Proceedings I.E.E.*, Paper No. 1177 R, September, 1951 (**98**, Part III, p. 383).

(4) SHANNON, C. E.: "Prediction and Entropy of Printed English," *Bell System Technical Journal*, 1951, **30**, p. 50.

(5) WOODWARD, P. M., and DAVIES, I. L.: "Information Theory and Inverse Probability in Telecommunication," *Proceedings I.E.E.*, Paper No. 1239 R, March, 1952 (**99**, Part III, p. 37).

(6) GOLDMAN, S.: "Information Theory" (Prentice Hall, 1953).

(7) GILBERT, E. N.: "A Comparison of Signalling Alphabets," *Bell System Technical Journal*, 1952, **31**, p. 504.

(8) BLACK, H. S.: "Modulation Theory" (van Nostrand, 1953).

(9) GOLDMAN, S.: "Frequency Analysis, Modulation and Noise" (McGraw-Hill, 1948).

(10) TREMELLEN, K. W., and COX, J. W.: "The Influence of Wave-Propagation on the Planning of Short-Wave Communication," *Journal I.E.E.*, 1947, **94**, Part IIIA, p. 200.

(11) THOMAS, H. A.: "Some Measurements of Atmospheric Noise at High Frequencies," *Proceedings I.E.E.*, Paper No. 1007 R, September, 1950 (**97**, Part III, p. 335).

(12) RECOMMENDATION No. 94: "Noise and Sensitivity of Receivers" (C.C.I.R. Documents of the VIIth Plenary Assembly, London, 1953).

(13) RYLE, M.: "Radio Astronomy," *The Physical Society Reports on Progress in Physics*, 1950, **13**, p. 184.

(14) WHITEHEAD, S.: "A Tentative Statistical Study of Domestic Radio Interference," *Journal I.E.E.*, 1943, **90**, Part III, p. 181.

(15) FUBINI, E. G., and JOHNSON, D. C.: "Signal-to-Noise Ratio in A.M. Receivers," *Proceedings of the Institute of Radio Engineers*, 1948, **36**, p. 1461.

(16) WEIGHTON, D.: "Impulsive Interference in Amplitude Modulation Receivers," *Journal I.E.E.*, 1948, **95**, Part III, p. 69.

(17) RECOMMENDATION No. 99: "Bandwidths and Signal-to-Noise Ratios in Complete Systems" (C.C.I.R. Documents of the VIIth Plenary Assembly, London, 1953).

(18) FELDMAN, C. B., and BENNET, W. R.: "Bandwidth and Transmission Performance," *Bell System Technical Journal*, 1949, **28**, p. 490.

(19) LANDON, V. F.: "Theoretical Analysis of Various Systems of Multiplex Transmission," *RCA Review*, 1948, **9**, pp. 287 and 433.

(20) SUNDE, E.: "Theoretical Fundamentals of Pulse Transmission," *Bell System Technical Journal*, 1954, **33**, pp. 721 and 987.

(21) CHERRY, E. C.: "The Transmission Characteristics of Asymmetric-Sideband Communication Networks," *Journal I.E.E.*, 1942, **89**, Part III, p. 19, and 1943, **90**, Part III, p. 75.

(22) CAWTHRA, W. A., and THOMSON, W. E.: "Bandwidth of a Sinusoidal Carrier Wave, Frequency-Modulated by a Rectangular Wave with Half-Sine-Wave Build-up," *Proceedings I.E.E.*, Paper No. 1066 R, January, 1951 (**98**, Part III, p. 69).

(23) SALINGER, H.: "Transients in Frequency Modulation," *Proceedings of the Institute of Radio Engineers*, 1942, **30**, p. 378.

(24) WHEELER, H. A.: "The Interpretation of Amplitude and Phase Distortion in Terms of Paired Echoes," *ibid.*, 1939, **27**, p. 359.

(25) HARTLEY, R. V. L.: "Steady State Delay as Related to Aperiodic Signals," *Bell System Technical Journal*, 1941, **20**, p. 222.

(26) ASSADOURIAN, F.: "Distortion of a Frequency-Modulated Signal by Small Loss and Phase Variations," *Proceedings of the Institute of Radio Engineers*, 1952, **40**, p. 172.

(27) MEDHURST, R. G.: "Harmonic Distortion of Frequency-Modulated Waves by Linear Networks," *Proceedings I.E.E.*, Paper No. 1650 R, May, 1954 (**101**, Part III, p. 171).

(28) LAMPARD, D. G.: "Harmonic and Intermodulation Distortion in 'Power Law' Devices," *Proceedings I.E.E.*, Monograph No. 55 R, December, 1952 (**100**, Part IV, p. 3).

(29) BROCKBANK, R. A., and WASS, C. A. A.: "Non-Linear Distortion in Transmission Systems," *Journal I.E.E.*, 1945, **92**, Part III, p. 45.

(30) WASS, C. A. A.: "A Table of Intermodulation Products," *ibid.*, 1948, **95**, III, p. 31.

(31) LEWIN, L.: "Interference in Multichannel Circuits," *Wireless Engineer*, 1950, **27**, p. 294.

(32) STARR, A. T., and WALKER, T. H.: "Microwave Radio Links," *Proceedings I.E.E.*, Paper No. 1290 R, March, 1952 (**99**, Part III, p. 241).

(33) BODE, H. W.: "Network Analysis and Feedback Amplifier Design" (van Nostrand, 1945).

(34) *GRAYSON, H., MCLEOD, T. S., DUNKLEY, R. A. G., and DAWSON, G.: "Circuit Technique in Frequency-Modulated Microwave Links," *Proceedings I.E.E.*, Paper No. 1291 R, March, 1952 (**99**, Part III, p. 256).

(35) LICKLIDER, J. C. R., and POLLACK, I.: "Effects of Differentiation, Integration and Infinite Peak Clipping upon the Intelligibility of Speech," *Journal of the Acoustical Society of America*, 1948, **20**, p. 42.

(36) *COLE, A. W., and SMALE, J. A.: "The Transmission of Pictures by Radio," *Proceedings I.E.E.*, Paper No. 1356 R, May, 1952 (**99**, Part III, p. 325).

(37) DRAFT REPORT OF SUB-GROUP XI-D: "Circuit Characteristics for the Transmission of Television Signals over Long Distances" (C.C.I.R. Documents of the VIIth Plenary Assembly, London, 1953).

(38) EASTMAN, A. V., and HORLE, L. C. F.: "The Generation of Spurious Signals by Non-Linearity of the Transmission Path," *Proceedings of the Institute of Radio Engineers*, 1940, **28**, p. 438.
(39) BLAKE, K. W.: "External Cross-Modulation in the 100 Mc/s Band," *Journal I.E.E.*, 1947, **94**, Part IIIA, p. 659.
(40) BABCOCK, W. C.: "Intermodulation Interference in Radio Systems," *Bell System Technical Journal*, 1953, **32**, p. 63.
(41) HUXLEY, L. G. H., and RATCLIFFE, J. A.: "A Survey of Ionospheric Cross-Modulation," *Proceedings I.E.E.*, Paper No. 876 R, September, 1949 (**96**, Part III, p. 433).
(42) *GLUYAS, T. M.: "Television Transmitter Considerations in Color Broadcasting," *RCA Review*, 1954, **15**, p. 312.
(43) KITCHEN, F. A., PRESSEY, B. G., and TREMELLEN, K. W.: "A Review of Present Knowledge of the Ionospheric Propagation of Very-Low, Low- and Medium-Frequency Waves," *Proceedings I.E.E.*, Paper No. 1463 R, March, 1953 (**100**, Part III, p. 100).
(44) PIGGOT, W. R.: "The Reflection and Absorption of Radio Waves in the Ionosphere," *Proceedings I.E.E.*, Paper No. 1434 R, March, 1953 (**100**, Part III, p. 61).
(45) SAXTON, J. A.: "Basic Ground-Wave Propagation Characteristics in the Frequency Band 50–800 Mc/s," *Proceedings I.E.E.*, Paper No. 1601 R, January, 1954 (**101**, Part III, p. 211).
(46) SAXTON, J. A., and HOPKINS, H. G.: "Some Adverse Influences of Meterological Factors on Marine Navigational Radar," *Proceedings I.E.E.*, Paper No. 1060 R, January, 1951 (**98**, Part III, p. 26).
(47) BRAY, W. J., LILLICRAP, H. G., and OWEN, F. C.: "The Fading Machine and its Use for the Investigation of the Effects of Frequency Selective Fading," *Journal I.E.E.*, 1947, **94**, Part IIIA, p. 283.
48) *FRIIS, H. T., and FELDMAN, C. B.: "A Multiple Unit Steerable Antenna for Short-Wave Reception," *Proceedings of the Institute of Radio Engineers*, 1937, **25**, p. 841.
(49) BRAMLEY, E. N.: "Diversity Effects in Spaced-Aerial Reception of Ionospheric Waves," *Proceedings I.E.E.*, Paper No. 1062 R, January, 1951 (**98**, Part III, p. 19).
(50) VAN WAMBECK, S. H., and ROSS, A. H.: "Performance of Diversity Receiving Systems," *Proceedings of the Institute of Radio Engineers*, 1951, **39**, p. 256.
(51) RECOMMENDATION No. 105: "Fading Allowances for the Various Classes of Service" (C.C.I.R. Documents of the VIIth Plenary Assembly, London, 1953).
(52) REPORT No. 27: "Fading of H.F. and M.F. Signals propagated by the Ionosphere" (C.C.I.R. Documents of the VIIth Plenary Assembly, London, 1953).
(53) CRAWFORD, A. B., and JAKES, W. C.: "Selective Fading of Microwaves," *Bell System Technical Journal*, 1952, **31**, p. 68.
(54) KAYLOR, R. L.: "A Statistical Study of Selective Fading of Super-High Frequency Radio Signals," *ibid.*, 1953, **32**, p. 1187.
(55) BOGLE, A. C.: "Some Aspects of Microwave Fading on an Optical Path over Sea," *Proceedings I.E.E.*, Paper No. 1373 R, September, 1952 (**99**, Part III, p. 236).
(56) BRAY, W. J., HOPKINS, H. G., KITCHEN, F. A., and SAXTON, J. A.: "Review of Long-Distance Radio-Wave Propagation above 30 Mc/s," *Proceedings I.E.E.*, Paper No. 1782 R, January, 1955 (**102** B, p. 87).
(57) BOOKER, H. G., and DE BETTENCOURT, J. T.: "Theory of Radio Transmission by Tropospheric Scattering using Very Narrow Beams," *Proceedings of the Institute of Radio Engineers*, 1955, **43**, p. 281.
(58) RECOMMENDATION No. 87: "Bandwidth of Emissions" (C.C.I.R. Documents of the VIIth Plenary Assembly, London, 1953).
(59) BOOTH, C. F.: "The Evolution of Frequency Control," *Proceedings I.E.E.*, 1951, **98**, Part III, p. 1.
(60) *RECOMMENDATION No. 96: "Frequency Stability of Receivers" (C.C.I.R. Documents of the VIIth Plenary Assembly, London, 1953).
(61) *RECOMMENDATION No. 95: "Selectivity of Receivers" (C.C.I.R. Documents of the VIIth Plenary Assembly, London, 1953).
(62) HAMMING, R. W.: "Error Detecting and Error Correcting Codes," *Bell System Technical Journal*, 1950, **29**, p. 147.
(63) *HAYTON, T., HUGHES, C. J., and SAUNDERS, R. L.: "Telegraph Codes and Code Converters," *Proceedings I.E.E.*, Paper No. 1585 R, November, 1953 (**101**, Part III, p. 137).
(64) *SOWTON, C. W., and BALCHIN, D. B.: "The New London Radio-Telephony Terminal," *Post Office Electrical Engineers' Journal*, April, 1951, **44**, p. 25.
(65) *LAWTON, J.: "The Reduction of Cross Talk on Trunk Circuits by the Use of the Volume Range Compressor and Expander," *ibid.*, April, 1939, **32**, p. 32.
(66) MALLINCKRODT, C. O.: "Instantaneous Compandors," *Bell System Technical Journal*, 1951, **30**, p. 706.
(67) LOZIER, J. C.: "Instantaneous Compandors on Narrow Band Speech Channels," *ibid.*, p. 1214.
(68) HALSEY, R. J., and SWAFFIELD, J.: "Analysis-Synthesis Telephony, with Special Reference to the Vocoder," *Journal I.E.E.*, 1948, **95**, Part III, p. 391.
(69) GABOR, D.: "New Possibilities in Speech Transmission," *Journal I.E.E.*, 1947, **94**, Part III, p. 369.
(70) MERTZ, P., and GRAY, F.: "Theory of Scanning," *Bell System Technical Journal*, 1934, **13**, p. 464.
(71) ABRAHAMS, I. C.: "The 'Frequency Interleaving' Principle in the N.T.S.C. Standards," *Proceedings of the Institute of Radio Engineers*, 1954, **42**, p. 81.
(72) CHERRY, E. C., and GOURIET, G. G.: "Some Possibilities for the Compression of Television Signals by Recoding," *Proceedings I.E.E.*, Paper No. 1401 R, January, 1953 (**100**, Part III, p. 9).
(73) KRETZMER, E. R.: "Statistics of Television Signals," *Bell System Technical Journal*, 1952, **31**, p. 751.
(74) HARRISON, C. W.: "Experiments with Linear Prediction in Television," *ibid.*, p. 764.
(75) *MUMFORD, A. H.: "Long-Distance Point-to-Point Communication," *Journal I.E.E.*, 1947, **94**, Part IIIA, p. 23.
(76) *BRAY, W. J., LILLICRAP, H. G., and LOWRY, W. R. H.: "The Design of Transmitter Drives and Receivers for Single-Sideband Systems," *Journal I.E.E.*, 1947, **94**, Part IIIA, p. 298.
(77) HARTLEY, R. V. L.: U.S. Patent 1666206, April, 1928.
(78) DARLINGTON, S.: "Realization of a Constant Phase Difference," *Bell System Technical Journal*, 1950, **29**, p. 94.
(79) *BARNES, G. W.: "A Single-Sideband Controlled-Carrier System for Aircraft Communication," *Proceedings I.E.E.*, Paper No. 1535 R, August, 1953 (**101**, Part III, p. 121).
(80) LAWSON, J. L., and UHLENBECK, G. E.: "Threshold Signals" (McGraw-Hill, 1950).
(81) SMITH, R. A.: "The Relative Advantages of Coherent and Incoherent Detectors: A Study of their Output Noise Spectra under Various Conditions," *Proceedings I.E.E.*, Monograph No. 6 R, August, 1951 (**98**, Part IV, p. 43).
(82) LEE, Y. W., CHEATHAM, T. P., and WIESNER, J. B.: "Application of Correlation Analysis to the Detection of Periodic Signals in Noise," *Proceedings of the Institute of Radio Engineers*, 1950, **38**, p. 1165.
(83) STUMPERS, F. L. H. M.: "Theory of Frequency-Modulated Noise," *ibid.*, 1948, **36**, p. 1081.
(84) MIDDLETON, D.: "On Theoretical Signal-to-Noise Ratio in F.M. Receivers," *Journal of Applied Physics*, 1949, **20**, p. 334.
(85) BRADLEY, W. E., and SMITH, D. B.: "Theory of Impulse Noise in Ideal Frequency-Modulation Receivers," *Proceedings of the Institute of Radio Engineers*, 1946, **34**, p. 743.
(86) *PETERSON, H. O., *et al.*: "Observations and Comparisons on Radio Telegraph Signalling by Frequency Shift and On-Off Keying," *RCA Review*, 1946, **7**, p. 11.
(87) RUDDLESDEN, R., FORSTER, E., and JELONEK, Z.: "Carrier-Frequency-Shift Telegraphy," *Journal I.E.E.*, 1947, **94**, Part IIIA, p. 379.
(88) BLACHMAN, N. M.: "A Comparison of the Informational Capacities of Amplitude- and Phase-Modulation Communication Systems," *Proceedings of the Institute of Radio Engineers*, 1953, **41**, p. 748.
(89) MONTGOMERY, G. F.: "A Comparison of Amplitude and Angle Modulation for Narrow-Band Communication of Binary-Coded Messages in Fluctuation Noise," *Proceedings of the Institute of Radio Engineers*, 1954, **42**, p. 447.
(90) MONTGOMERY, G. F.: "Message Error in Diversity Frequency-Shift Reception," *ibid.*, p. 1184.
(91) *MERRIMAN, J. H. H., HAMER, R., and HORSFIELD, B. R.: "A Subscriber's Battery-Operated Single-Channel V.H.F. Radio-Telephone Equipment," *Post Office Electrical Engineers' Journal*, 1951, **44**, p. 75.
(92) *MERRIMAN, J. H. H., and WHITE, R. W.: "The Application of Frequency Modulation to V.H.F. Multi-Channel Telephony," *Journal I.E.E.*, 1947, **94**, Part IIIA, p. 649.
(93) COOK, D., JELONEK, Z., OXFORD, A. J., and FITCH, E.: "Pulse Communication," *ibid.*, p. 83.
(94) CHALK, J. H. H.: "The Optimum Pulse-Shape for Pulse Communication," *Proceedings I.E.E.*, Paper No. 904 R, March, 1950 (**97**, Part III, p. 88).
(95) OLIVER, B. M., PIERCE, J. R., and SHANNON, C. E.: "The Philosophy of P.C.M.," *Proceedings of the Institute of Radio Engineers*, 1948, **36**, p. 1324.
(96) SCHOUREN, J. F., DE JAGER, F., and GREEFKES, J. A.: "Delta-Modulation, A New Modulation System for Telecommunication," *Philips Technical Review*, 1952, **13**, p. 237.
(97) BENNET, W. R.: "Time-Division Multiplex Systems," *Bell System Technical Journal*, 1941, **20**, p. 199.
(98) MOSKOWITZ, S., DIVEN, L., and FEIT, L.: "Cross-Talk Considerations in Time-Division Multiplex Systems," *Proceedings of the Institute of Radio Engineers*, 1950, **38**, p. 1330.
(99) HOPPER, A. L.: "Non-Synchronous Pulse Multiplex System," *Electronics*, 1952, **28**, p. 116.
(100) *BUFF, C.: "Twinplex and Twinmode Radiotelegraph Systems," *Electrical Communication*, 1952, p. 20.

Paper No. 1907 M
Nov. 1955

621.3.018 : 621—52

AN APPROXIMATE METHOD FOR OBTAINING TRANSIENT RESPONSE FROM FREQUENCY RESPONSE

By H. H. ROSENBROCK, B.Sc.(Eng.), Ph.D., Associate Member.

(*The paper was first received* 18*th March, and in revised form* 8*th June*, 1955.)

SUMMARY

A number of approximate methods are known for obtaining the transient response of a linear system from its frequency response when the latter is known numerically but not as an analytical function. The paper presents a new method for performing this transformation which is believed to be considerably quicker and more convenient than previous methods, and to offer a greater physical insight into the behaviour of the system.

The method can be used for the inverse problem of finding the frequency response of a system from its measured transient response. It also offers a means for obtaining an analytical function which represents with known accuracy the behaviour of a system of which either the frequency response or the transient response has been measured.

(1) INTRODUCTION

If the transfer function of a stable linear system, initially at rest, is F(p), then the response f(t) of the system to a unit step-function is[1]

$$f(t) = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} \frac{F(p)}{p}\varepsilon^{pt}dp \quad . \quad . \quad . \quad . \quad (1)$$

This formula holds when F(p)/p has poles on the imaginary axis, provided that the path of integration is taken as in Fig. 1. If

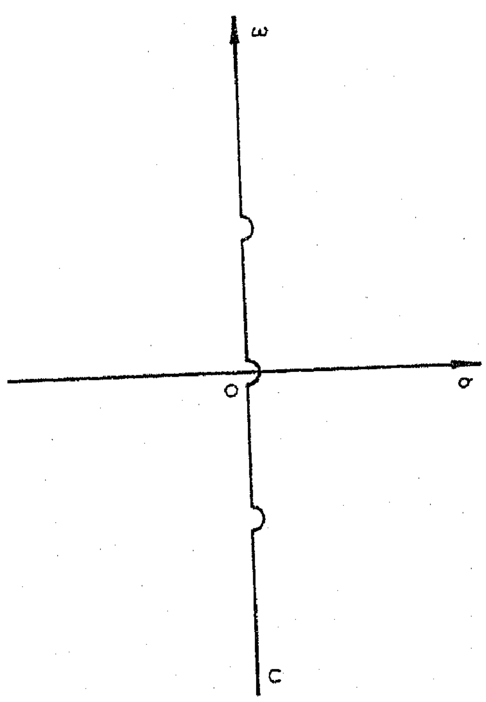


Fig. 1.—Path of integration in the complex plane of $p = \sigma + j\omega$.

The path of integration, C, is indented into the right-hand half-plane to avoid any poles on the imaginary axis.

F(p) is known and has a simple form, eqn. (1) may be evaluated analytically, and for more complicated functions F(p) the evaluation may be performed by means of an analogue computer[2,3] or by a graphical construction.[4,5]

It may happen, however, that F(p) is known only by experiment from measurements taken with harmonic inputs; i.e. we know F($j\omega$) as a complex function of ω. This situation will often arise in the control of processes which are not readily analysed mathematically. We then have the choice of approximating to F($j\omega$) by means of an analytical function, which may not be easy, or of using one of several graphical methods which have been developed.[6,7,8,9,10]

In the following Section a new method for approximating to the solution of eqn. (1) is presented. It is believed to be considerably quicker and more convenient than existing methods, and to offer a greater physical insight into the behaviour of the system. It makes use of a specially-constructed transparent cursor and can be used with measured functions.

(2) APPROXIMATE METHOD FOR OBTAINING THE TRANSIENT RESPONSE

We shall be concerned with functions F(p) which have no poles in the right-hand half-plane or on the imaginary axis, the only possible pole of F(p)/p being at the origin. It is shown in Section 12 that the two following equations can then be derived from eqn. (1).

$$f(t) = \frac{2}{\pi}\int_{\omega=0}^{\infty} [\mathscr{R}F(j\omega)] \sin \omega t \, d(\log \omega) \quad . \quad . \quad (2)$$

$$f(t) - f(\infty) = \frac{2}{\pi}\int_{\omega=0}^{\infty} [\mathscr{I}F(j\omega)] \cos \omega t \, d(\log \omega) \quad . \quad (3)$$

If F(p)/p has no pole at the origin, f(∞) = 0.

Since we have

$$f(\infty) = \lim_{\omega \to 0} F(j\omega) \quad . \quad . \quad . \quad . \quad . \quad (4)$$

which is readily found, it is possible to use either eqn. (2) or eqn. (3) to find f(t). For several reasons, however, eqn. (3) is the more convenient. The reasons are given here, though they will become clearer when the approximate method has been considered.

(*a*) $\mathscr{I}$F($j\omega$) tends to zero as ω tends to zero or to infinity. This simplifies the method explained below for evaluating eqn. (1).

(*b*) When $\mathscr{I}$F($j\omega$) is plotted against log ω the net area under the curve is finite and gives f(0) − f(∞). This forms a convenient check.

(*c*) In the application to automatic control the function F($j\omega$) will represent either $1/(1 + G)$ or $G/(1 + G)$, where G is the open-loop harmonic response of the servo system. Since $1/(1 + G) = 1 - G/(1 + G)$ it follows that $\mathscr{I}1/(1 + G) = -\mathscr{I}G/(1 + G) = -\mathscr{I}1/(1 + G^{-1}) = \mathscr{I}G^{-1}/(1 + G^{-1})$. We may therefore work with G or with G^{-1} indifferently.

(*d*) Since the important harmonic terms in f(t) normally have the phase of cos ωt, eqn. (3) gives a more natural understanding of the behaviour of the system. For example, if the system gives a single damped oscillation with frequency ω_0, f(t) will reach its greatest overshoot where $\omega_0 t \simeq \pi$. At this point cos $\omega_0 t$ will be nearly one, and the contribution of F($j\omega_0$) to the integral in eqn. (3) will be large. On the other hand, sin $\omega_0 t$ will be small, and F($j\omega_0$) will not have much effect on the integral of eqn. (2).

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
Dr. Rosenbrock is with Costain-John Brown, Ltd.

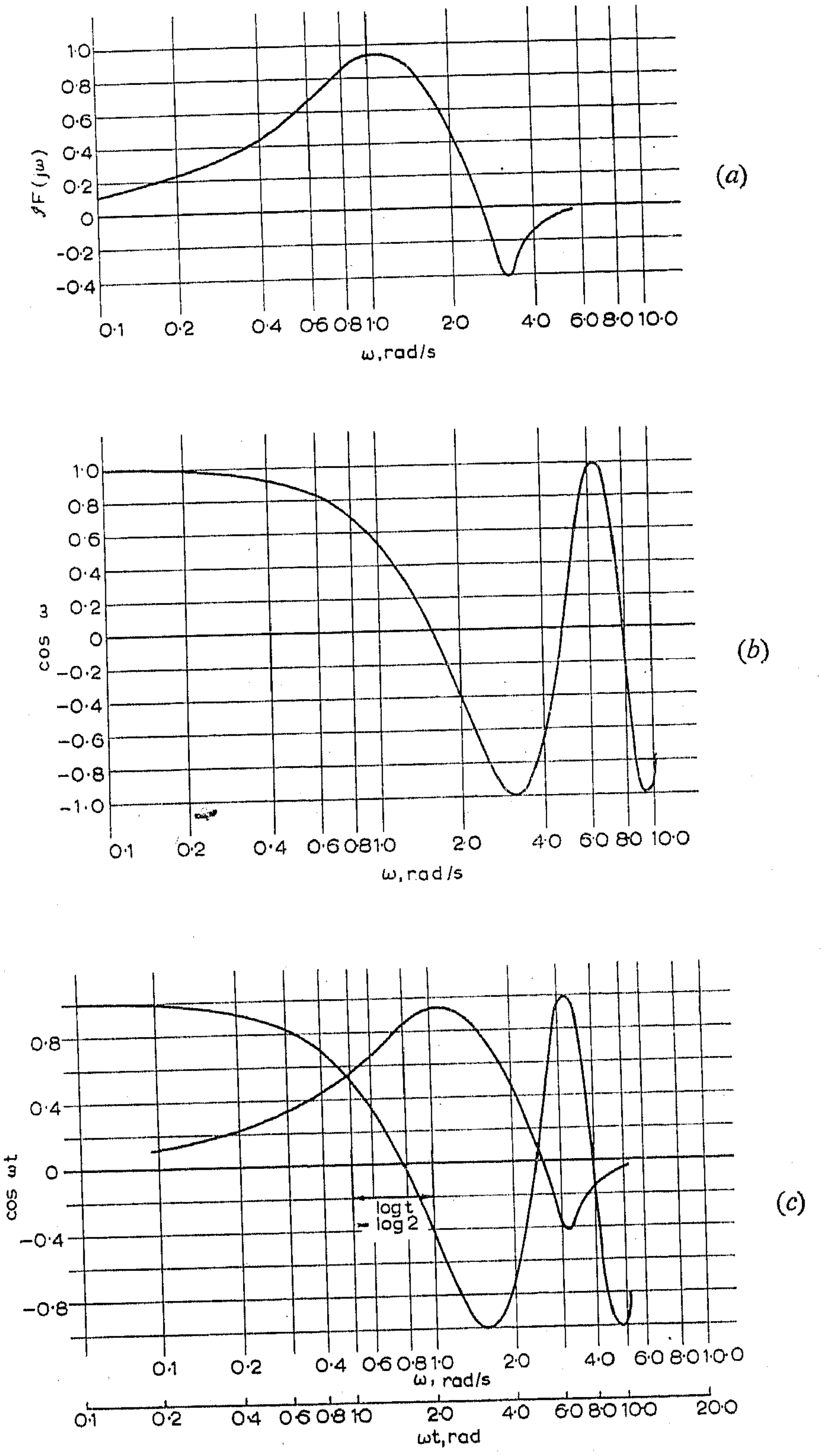


Fig. 2.—Derivation of the cursor.

(*a*) The imaginary part of a function F(*j*ω) plotted against log ω.
(*b*) The graph of cos ω plotted against log ω.
(*c*) When (*b*) is moved a distance log *t* to the left and superimposed on (*a*) it shows at each point the value of cos ω*t*.

Considering eqn. (3), let us first plot $\mathscr{I}F(j\omega)$ against $\log \omega$, as in Fig. 2(*a*). On a transparent cursor let us plot $\cos \omega$ against $\log \omega$, using the same scales as before. This is shown in Fig. 2(*b*). If the cursor is placed over the graph of $\mathscr{I}F(j\omega)$ and displaced a distance $\log t$ to the left, as in Fig. 2(*c*), it gives for each value of $\mathscr{I}F(j\omega)$ the value of $\cos \omega t$. Multiplying these two values and finding the area under the resulting curve to a suitable scale would, from eqn. (3), give $f(t) - f(\infty)$.

A more convenient method is available, however.[16] We produce a transparent cursor (see Fig. 3) with ordinates each in the centre of one of a number of strips into which we divide the graph of $\cos \omega$ against $\log \omega$. On each ordinate is marked a scale with intervals Δ inversely proportional to the product of height and width of its strip. Then placing the cursor over the graph of $\mathscr{I}F(j\omega)$ we add the values of $\mathscr{I}F(j\omega)$ indicated at each

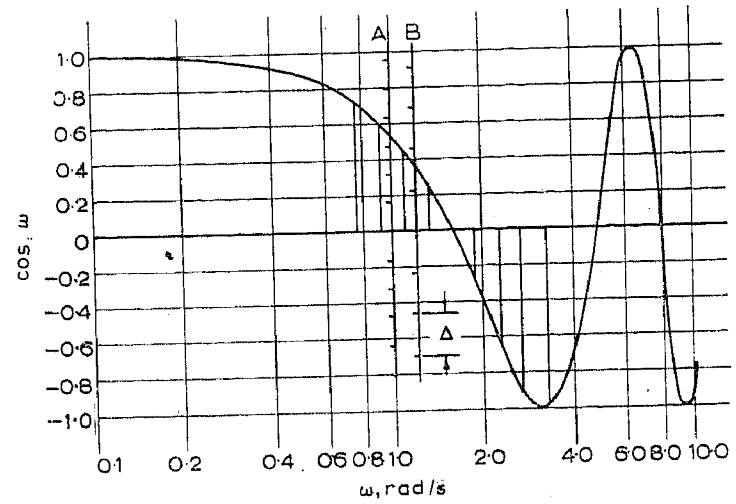


Fig. 3.—Derivation of the cursor.

The graph of cos ω is divided into strips, and ordinates A, B, etc., are erected in the middle of each strip.

Table 1

NUMERICAL VALUES FOR CURSOR

Ordinate	ω	Δ
Grid	Spaced as ordinates 1 to 21	0·0858
1	0·110	0·086
2	0·133	0·087
3	0·159	0·087
4	0·191	0·087
5	0·230	0·088
6	0·276	0·089
7	0·331	0·091
8	0·398	0·093
9	0·478	0·097
10	0·574	0·102
11	0·689	0·111
12	0·828	0·127
13	0·994	0·157
14	1·19	0·233
15	1·43	0·626
16	1·72	0·560
17	2·07	0·180
18	2·48	0·109
19	2·98	0·088
20	3·58	0·096
21	4·30	0·220
22	5·13	0·200
23	6·08	0·110
24	7·21	0·155
25	8·54	0·160
26	10·1	0·130
27	12·5	0·098
28	15·6	0·123
29	18·8	0·148
30	21·9	0·173
31	25·1	0·197
32	28·2	0·222
33	31·4	0·247
34	34·5	0·271
35	37·7	0·568

successive ordinate, taking due account of sign. We thus approximate to the integral in eqn. (3).

One further refinement can be made. We are approximating to the integral in eqn. (3) by a sum of the form

$$\sum_r \mathscr{I}\mathrm{F}(j\omega_r)\cos\omega_r t \times \delta_r(\log\omega) \quad . \quad . \quad . \quad (5)$$

and the approximation will be good if neither $\mathscr{I}\mathrm{F}(j\omega)$ nor $\cos\omega t$ changes rapidly in the range $\delta_r(\log\omega)$. The functions $\mathscr{I}\mathrm{F}(j\omega)$ with which we shall deal will usually fulfil this condition; but Fig. 2(*b*) shows that $\cos\omega t$ does not do so when ωt is large. It is then better to replace the factor $\cos\omega_r t \times \delta_r(\log\omega)$ in eqn. (5) by the area under the curve $\cos\omega t$ in the range $\delta_r(\log\omega)$. This area can be obtained from the cosine-integral function $\mathrm{Ci}(\omega)$.

The values adopted for such a cursor are listed in Table 1, and the cursor is shown in Fig. 4. For ordinates 1 to 21 the value size of the intervals Δ is so adjusted that the height of a unit step is 100. The value obtained by use of the cursor is thus the percentage departure of $\mathrm{f}(t)$ from its final steady value $\mathrm{f}(\infty)$. For a zero-order servo system, $|\mathrm{f}(t) - \mathrm{f}(\infty)|$ must approach $|K_p/(K_p+1)|$ as t tends to zero, where K_p is the gain for $\omega = 0$. For a servo system of the first or higher order $|\mathrm{f}(t) - \mathrm{f}(\infty)|$ will tend to one as t tends to zero. Using the grid on the left of the cursor we may easily check these values, so checking the graph of $\mathscr{I}\mathrm{F}(j\omega)$.

The signs of $\mathscr{I}\mathrm{F}(j\omega)$ and of $\mathrm{f}(t) - \mathrm{f}(\infty)$ may be ignored during the calculation and inserted at the end, since they are physically obvious. We shall therefore show $\mathscr{I}\mathrm{F}(j\omega)$ with its greater area above the base line, and the cursor is arranged accordingly. With this convention for the sign of $\mathscr{I}\mathrm{F}(j\omega)$, a negative result (i.e. an excess of readings on the red ordinates of the cursor) always indicates overshoot.

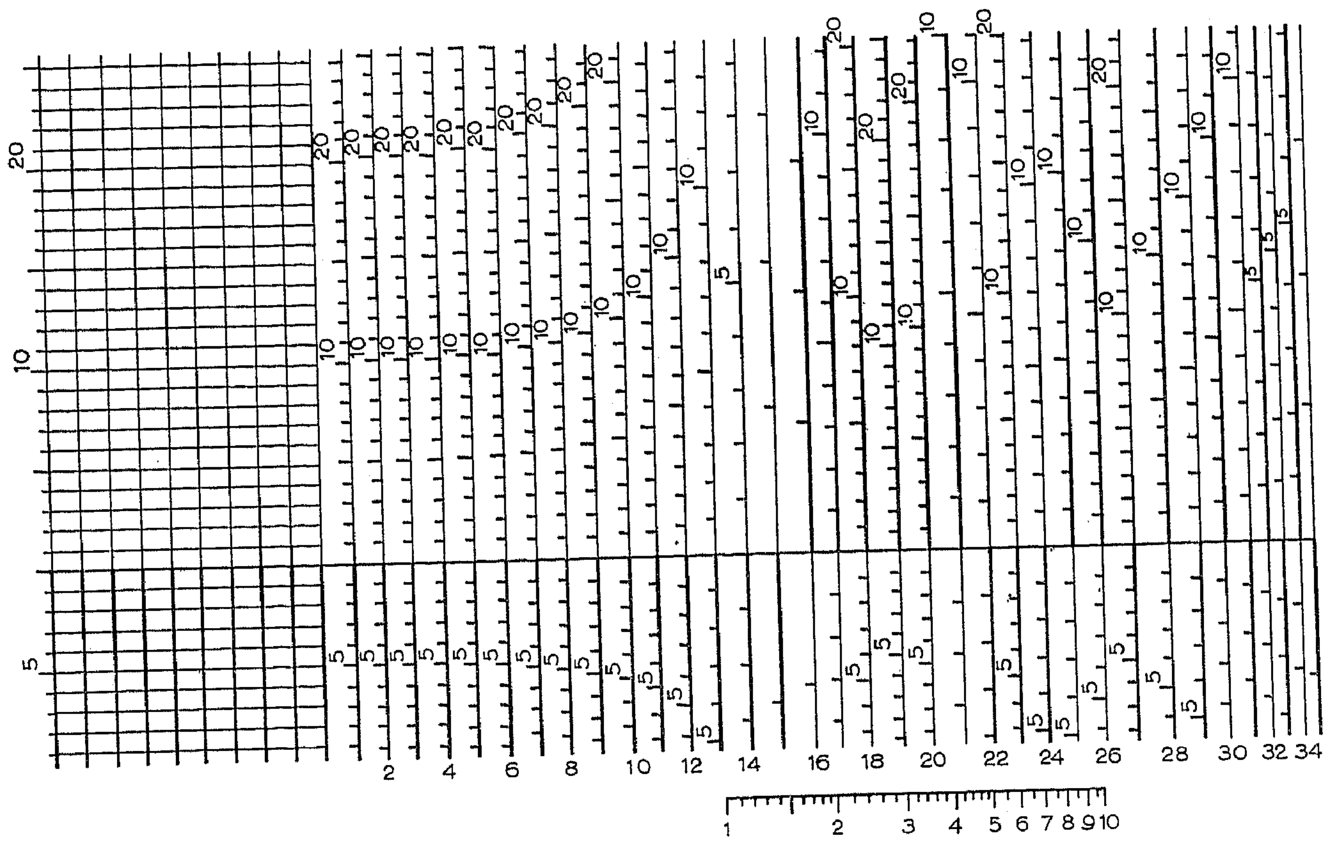


Fig. 4.—Diagram of the cursor.

The parts of the ordinates drawn more heavily are filled in the original with a red compound, and the remaining parts with black.

of Δ was obtained from the product of $\cos\omega t$ and $\delta(\log\omega)$. For ordinates 22 onwards, Δ was calculated from $\mathrm{Ci}(\omega)$. Ordinate 35 has a value of Δ corresponding to only a part of the area under the associated loop of $\cos\omega$; up to a value of ω, in fact, which makes $\mathrm{Ci}(\omega) = 0$. This reduces the error when the graph of $\mathscr{I}\mathrm{F}(j\omega)$ extends to the right of ordinate 35.

A grid is extended to the left of ordinate 1 with a constant spacing corresponding to $\cos\omega t = 1$, and is useful for checking that the graph of $\mathscr{I}\mathrm{F}(j\omega)$ has the correct area. A scale of t is provided, and is read against $\omega = 1$ on the graph of $\mathscr{I}\mathrm{F}(j\omega)$. If t is outside the limits, 1 to 10, within which this can be done directly, we read t against $\omega = 10^n$ and then multiply by 10^{-n}. For example, if $t = 2\cdot5$ is above $\omega = 10$, the true value of t is $0\cdot25$ sec. The cursor was prepared for use with a commercially available graph paper having one scale linear and the other logarithmic covering three "cycles" (i.e. decades).

In Fig. 4 parts of the ordinates are drawn more heavily. These after scribing on the cursor were filled with a red compound, the remaining parts being filled with black. In use, all the intercepts on red ordinates are added separately, and all those on black. The difference between the two sums gives $\mathrm{f}(t) - \mathrm{f}(\infty)$. The

(2.1) Methods for Obtaining $\mathscr{I}\mathrm{F}(j\omega)$

Since $\mathrm{F}(j\omega)$ will be $1/(1+G)$ or $G/(1+G) = 1/(1+G^{-1})$, and since these, apart from sign, have the same imaginary parts, we may use the locus of G or G^{-1} indifferently. Fig. 5 shows the curves of constant imaginary part in either case. They are circles of diameter $1/[\mathscr{I}\mathrm{F}(j\omega)]$ tangent to the real axis at $(-1, 0)$.

Figs. 2(*b*) and 5 give a simple means of visualizing the transient response corresponding to a given harmonic locus. For example, if a first-order servo system gives a function $\mathscr{I}\mathrm{F}(j\omega)$ with a single positive peak and no large negative part, and if the peak does not exceed a height of $0\cdot9$ (i.e. about 10 when measured on the grid at the left-hand side of the cursor), then the transient response must necessarily have only a small overshoot. For the net area under the graph of $\mathscr{I}\mathrm{F}(j\omega)$ must be 100, and the width must therefore be more than enough to cover 10 ordinates, and in practice probably much more. The greatest possible overshoot will be found by arranging the cursor to give the greatest negative value of the sum of intercepts. From Fig. 2(*b*), or directly from the cursor, it is evident that this cannot be large. When we know not only the peak value of $\mathscr{I}\mathrm{F}(j\omega)$ but also

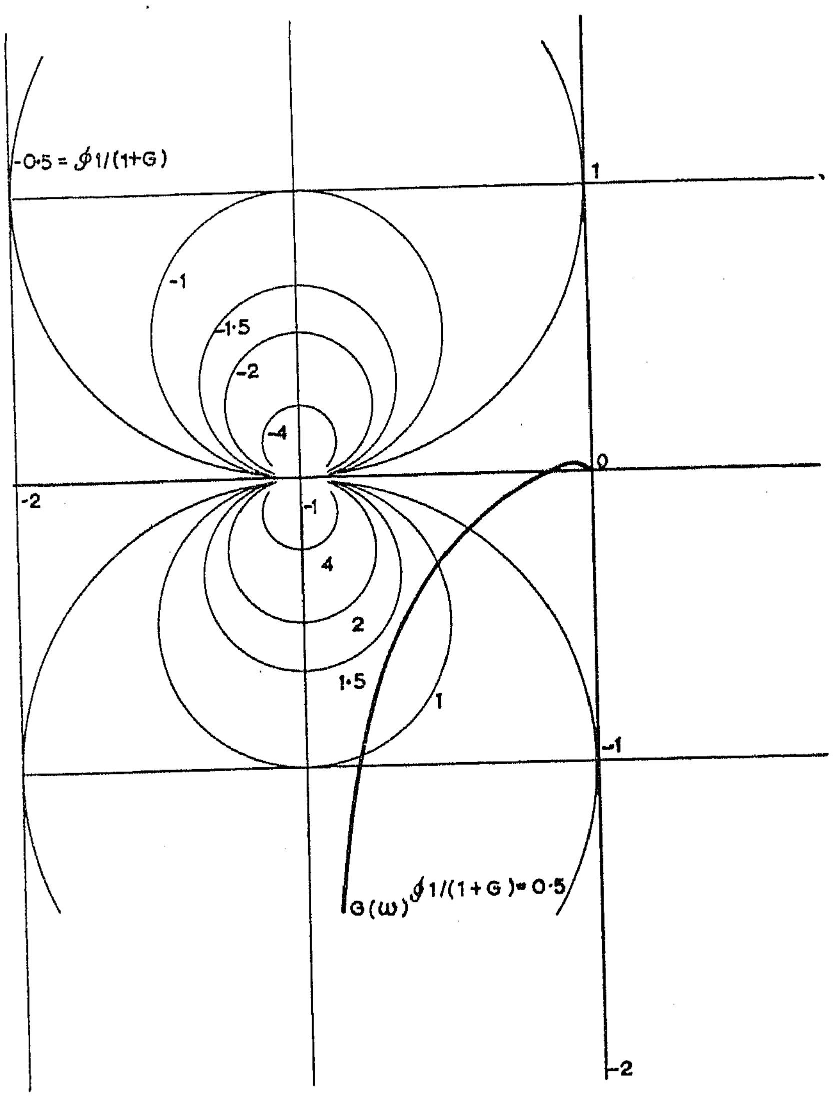


Fig. 5.—Curves of constant $\mathscr{I}1/(1 + G)$ on the plane of G.

The curves of constant $\mathscr{I}G/(1 + G)$ on the plane of G^{-1} are identical with those shown, as are curves of $-\mathscr{I}G/(1 + G)$ on the plane of G, or of $-\mathscr{I}1/(1 + G)$ on the plane of G^{-1}.

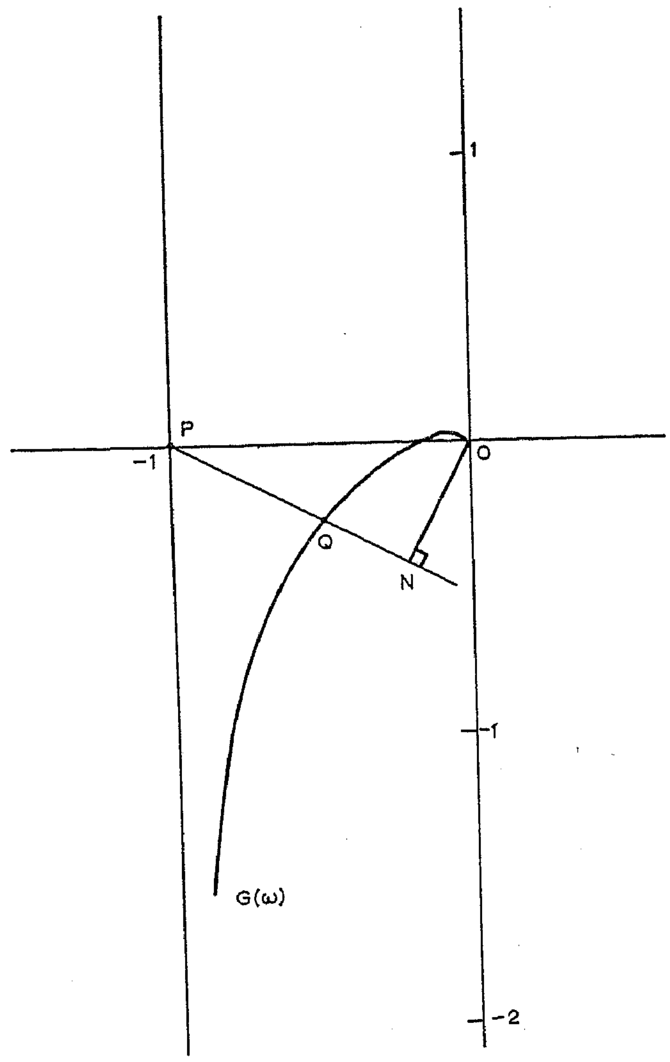


Fig. 6.—Construction for finding $\mathscr{I}1/(1 + G)$ or $\mathscr{I}G/(1 + G)$.

$\mathscr{I}1/(1 + G) = -\mathscr{I}G/(1 + G) = $ ON/PQ, positive if N is below the line OP. On the plane of G^{-1} we have a similar result but with reversed signs.

something about its shape, we may set closer limits to the transient response.

Although Fig. 5 is valuable in allowing us to visualize the transient response, it does not give the best means of finding $\mathscr{I}F(j\omega)$. The value of G or G^{-1} will usually be computed for only a few values of ω, and to use Fig. 5 we must either interpolate between these widely-spaced points on the locus or draw in a large number of circles and interpolate between them. A better method is shown in Fig. 6, where $\mathscr{I}F(j\omega)$ is given by ON/PQ.

An alternative construction is shown in Fig. 7. A set-square is arranged so that one side passes through the point $(-1, 0)$, and the right-angle lies on the locus. The intercept, on an ordinate through $(-1, 0)$, of the other side of the set-square gives $1/[\mathscr{I}F(j\omega)]$. This method is useful for finding the greatest value attained by $\mathscr{I}F(j\omega)$ at a peak.

Both of the constructions just given can be applied either to the locus of G or to that of G^{-1}. The appropriate sign will be evident in each case.

We may also find $\mathscr{I}F(j\omega)$ from a plot of $\log|G|$ against $\arg G$. A chart is given in Reference 7, p. 342, which gives $\mathscr{R}G/(1 + G)$ when $\log|G|$ and $\arg G$ are known. A similar chart can be prepared to give $\mathscr{I}G/(1 + G)$ or, what is the same thing apart from sign, $\mathscr{I}1/(1 + G)$.

(3) EXAMPLES

The first example is a position-control system comprising a Ward Leonard set driving an inertia load, the generator having a split field fed by a valve amplifier.[5] For a mechanical time-

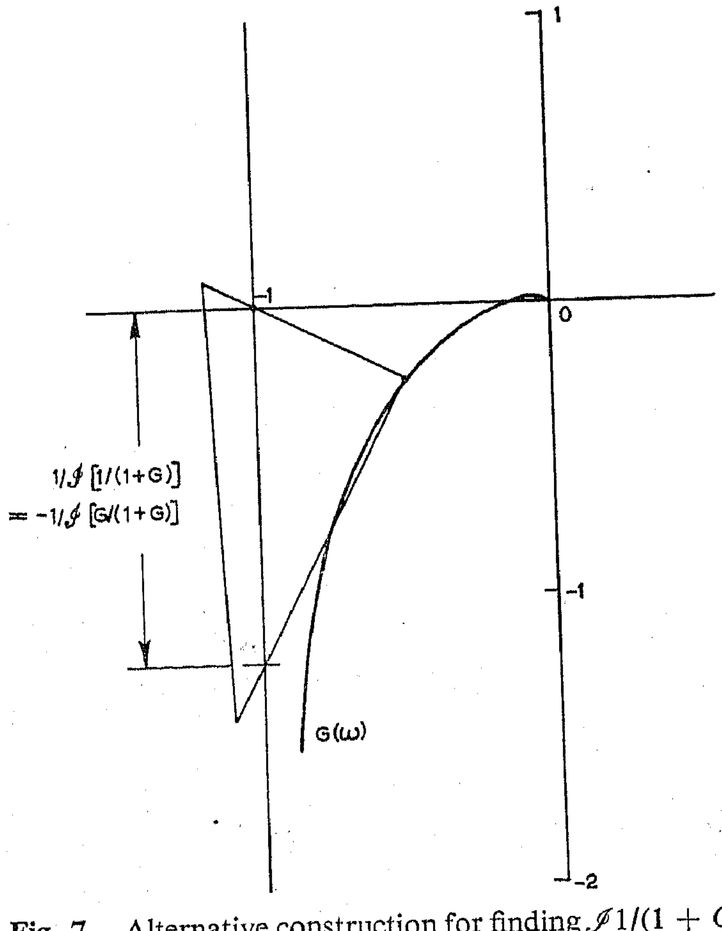


Fig. 7.—Alternative construction for finding $\mathscr{I}1/(1 + G)$ or $\mathscr{I}G/(1 + G)$.

On the plane of G^{-1} the signs are reversed.

constant of 5 sec, a field time-constant 1 sec, and a velocity-error constant of one-third we obtain

$$\epsilon(p) = \frac{1}{p[1 + G(p)]} = \frac{5p^2 + 6p + 1}{5p^3 + 6p^2 + p + 0\cdot333} \quad . \quad (6)$$

whence $$G^{-1}(\omega) = -j15\omega^3 - 18\omega^2 + 3j\omega \quad . \quad . \quad . \quad (7)$$

The last expression is easily calculated, and the corresponding locus is given in Fig. 8. By the methods of Section 2.1 we

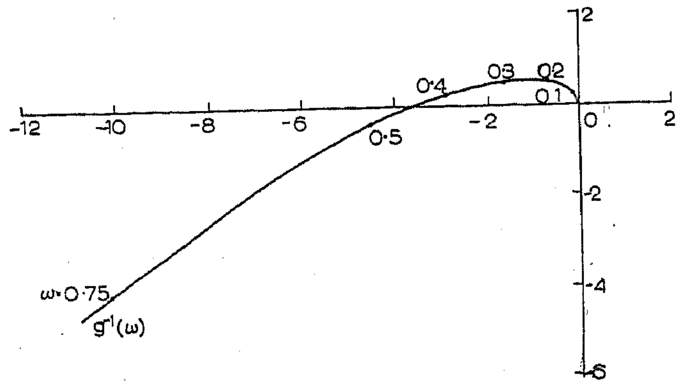


Fig. 8.—Plot of $G^{-1}(\omega)$ for a particular servo system.

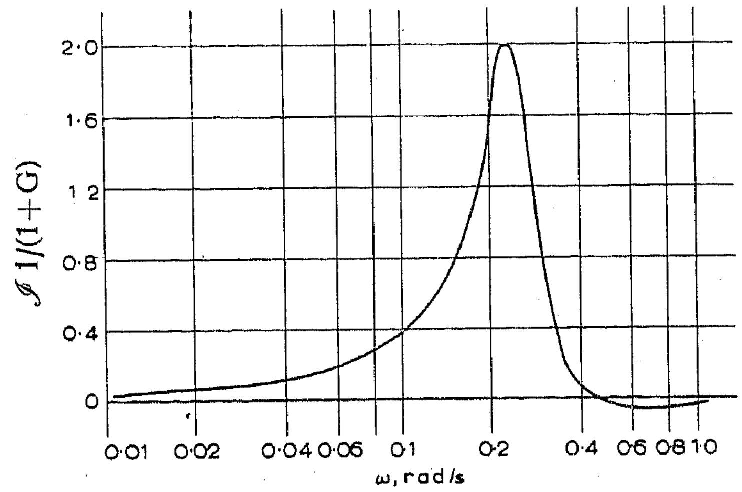


Fig. 9.—$\mathscr{I}\, 1/(1 + G)$ corresponding to $G^{-1}(\omega)$ shown in Fig. 8.

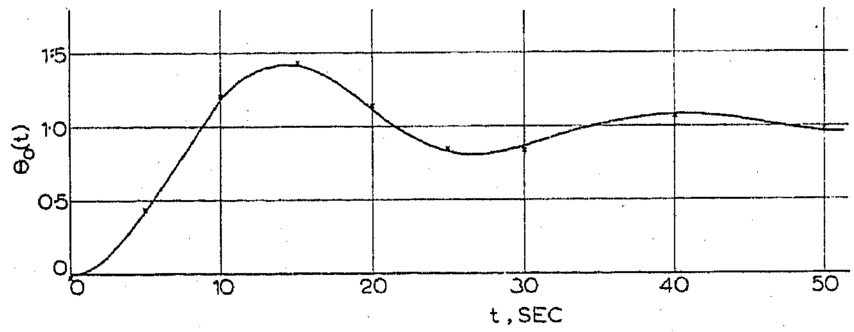


Fig. 10.—Comparison of calculated and approximate transients for the system characterized by Fig. 8.

The curve is based on calculated values, and the crosses show values obtained by means of the cursor.

derive the imaginary part of $1/(1 + G^{-1}) = G/(1 + G)$, which with its sign reversed is shown against a logarithmic scale of ω in Fig. 9. Use of the cursor then allows us to obtain with little effort the points marked with crosses in Fig. 10.

By expressing $\epsilon(p)$ in partial fractions and using the methods of the Laplace transformation, we find that the transient response is given by

$$\epsilon(t) = 0\cdot0613\varepsilon^{-1\cdot0714t} + \varepsilon^{-0\cdot0643t}\,(0\cdot9387\cos 0\cdot2410t + 0\cdot5093\sin 0\cdot2410t) \quad . \quad (8)$$

From this expression we obtain the response $\theta_0(t)$ shown by Fig. 10. Table 2 gives the values of $\theta_0(t)$ obtained by the two

Table 2

COMPARISON OF RESULTS

	$\theta_0(t)$	
	By Laplace transform	By approximate method
0	0	−0·01
5	0·41	0·43
10	1·19	1·21
15	1·41	1·43
20	1·11	1·14
25	0·84	0·85
30	0·86	0·84
40	1·08	1·07
50	0·98	0·99

methods, and in Table 3 are listed the readings obtained from the cursor for two values of t.

One point in Table 3 deserves attention. For small values of ω, $\mathscr{I}G/(1 + G)$ is proportional to ω. This fact may be used to

Table 3

DETAILED RESULTS FROM CURSOR

$t = 0$		$t = 15$	
Red	Black	Red	Black
0·5	0·2	0·6	0·2
0·7	2·0	0·1	0·4
0·5	8·0	4·3	0·3
0·3	21·0	20·4	0·4
	22·8	18·2	1·2
	12·9	8·8	0·5
	7·8	3·8	1·2
	6·1	0·9	1·3
	4·4		1·3
	3·3		1·2
	2·7		1·0
	2·0		5·0*
	1·7		
	1·3		
	6·5*		
−2·0	102·7	−57·1	14·0
	2	14·0	
	100·7	−43·1	

* Five times the last figure read from the cursor.

extend the graph of $\mathscr{I}G/(1 + G)$ without plotting G^{-1}, but it is not necessary to carry the graph down to very low values of ω. We notice that

$$\frac{200}{\pi}\int_{\omega=0}^{\omega_1} a\omega\, d(\log\omega) = \frac{200}{\pi}a\omega_1 \quad . \quad . \quad . \quad . \quad (9)$$

If we sum the readings on the cursor down to a value of ω at which $\cos\omega t$ is effectively one, the last reading will be $a\omega_1$ measured in units of $0\cdot0858$ (see Table 1). The area to the left of this ordinate is $200/\pi \times 0\cdot0858 \simeq 5\cdot5$ times the last reading. Subtracting one half the area of the strip corresponding to the last ordinate, which would otherwise be included twice, we find that we must add five times the last reading written down. This is shown in Table 3.

If the graph of $\mathscr{I}\,\mathrm{F}(j\omega)$ behaves at low frequencies like ω^n, we find in the same way that the necessary correction is $(5\cdot5/n - 0\cdot5)$ times the last reading written down.

The example just given represents a system which is under-damped by normal standards. The second example represents a system having very small overshoot, and is an example given by Floyd.[7] The open-loop transfer function is

$$G(p) = \frac{18\cdot72}{[(p+1)^2+1][(p+0\cdot6)^2+9] - 18\cdot72} \quad . \quad (10)$$

and again we may very easily calculate G^{-1}, which is shown in Fig. 11. Fig. 12 shows $\mathscr{I}\,1/(1+G)$ plotted against log ω, and Fig. 13 shows the points obtained by means of the cursor.

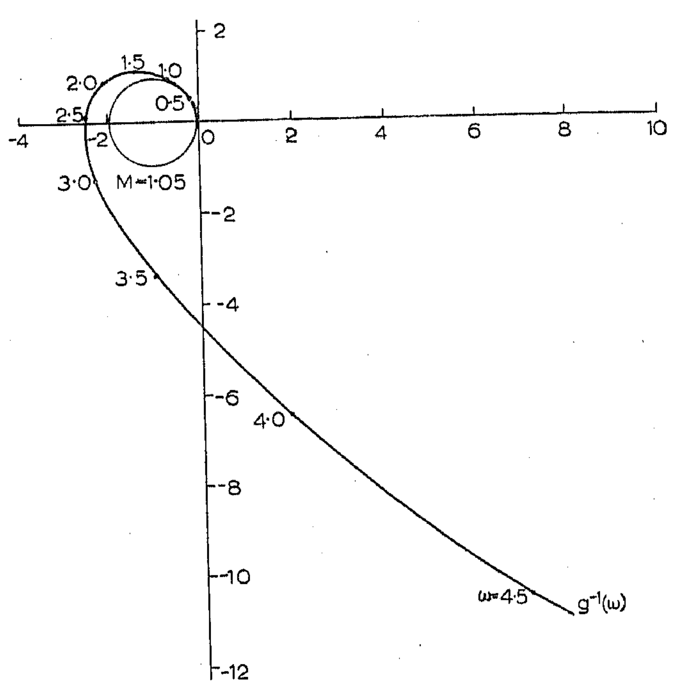


Fig. 11.—Plot of $G^{-1}(\omega)$ for a particular servo system.

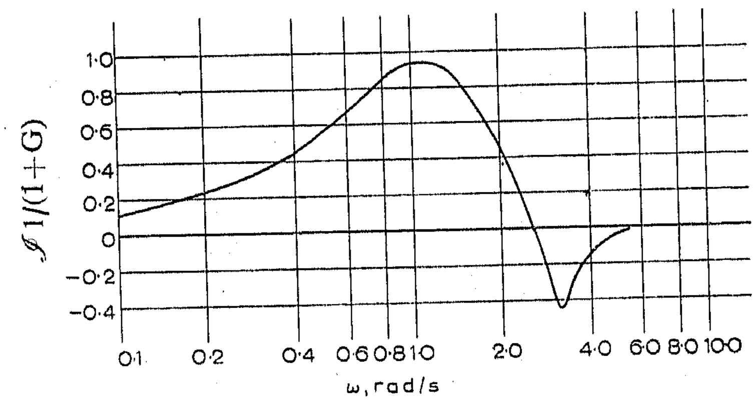


Fig. 12.—$\mathscr{I}\,1/(1+G)$ corresponding to $G^{-1}(\omega)$ shown in Fig. 11.

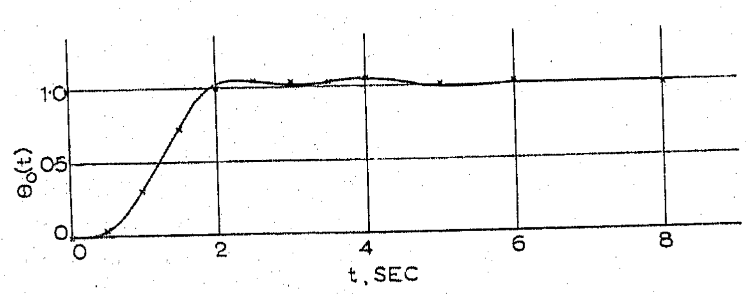


Fig. 13.—Comparison of calculated and approximate transients for the system characterized by Fig. 11.

The curve is based on calculated values, and the crosses show values obtained by means of the cursor.

The response of the system to a unit impulse is given analytically by Floyd. Integrating his result we obtain for the response to a unit step

$$\theta_0(t) = 1 - 1\cdot14\varepsilon^{-t}\,[\sin(t + 5\cdot6^\circ) + \cos(t + 5\cdot6^\circ)]$$
$$+ \varepsilon^{-0\cdot6t}[0\cdot0488\sin(3t + 17^\circ) + 0\cdot2439\cos(3t + 17^\circ)] \quad . \quad . \quad . \quad . \quad (11)$$

This function is shown by the graph in Fig. 13, and in Table 4 the values of $\theta_0(t)$ found by the two methods are compared.

Table 4

Comparison of Results

	$\theta_0(t)$	
	By Laplace transform	By approximate method
0	0	−0·01
0·5	0·03	0·03
1·0	0·31	0·30
1·5	0·74	0·73
2·0	1·02	0·99
2·5	1·04	1·04
3·0	1·02	1·03
3·5	1·03	1·03
4·0	1·05	1·05
5·0	0·99	1·00
6·0	1·00	1·02
8·0	1·00	0·99

(4) ACCURACY OF THE METHOD

It will be seen from Tables 2 and 4 that the error in $\theta_0(t)$ does not exceed 3%. Since no special care was taken with the graphical work, which was done on quarto or foolscap graph paper, this is about the order of accuracy which would be expected from the drawing alone. The cursor evidently produces only small errors; smaller in fact than one would expect from such a crude device. The reason is probably that where the contribution to f(t) is greatest, i.e. where $\mathscr{I}\,\mathrm{F}(j\omega)$ and $\cos\omega t$ are both large, they will usually be changing slowly, and the corresponding terms in expression (5) will have a small error. Where the proportional error in the terms of expression (5) is larger, the magnitude of the terms tends to be smaller.

(5) INTERPRETATION OF THE GRAPH OF $\mathscr{I}\,\mathrm{F}(j\omega)$

In Fig. 11 the circle for $M = 1\cdot05$ has been inserted, and it is evident that the servo-system resonance is very flat and occurs between $\omega = 0$ and $\omega = 1$. This tells us nothing useful about the poles of $1/(1 + G)$, and indeed is rather misleading. Fig. 12, on the other hand, suggests immediately a well-damped mode with $\omega \simeq 1$ and a second mode with smaller amplitude, opposed phase, and less damping, having $\omega \simeq 3$. Eqn. (11) shows that these features are in fact present.

We are therefore led to study the shapes of graph which can arise for the imaginary part of different functions $\mathrm{F}(j\omega)$. Fig. 14 shows the imaginary part of $\mathrm{F}(j\omega)$ for the function $\varepsilon^{-\alpha t}$, plotted against the logarithm of the dimensionless variable $u = \omega/\alpha$. Fig. 15 gives corresponding graphs for the function $\varepsilon^{-\zeta\omega_0 t}\cos\omega_0 t$ plotted against the logarithm of $u = \omega/\omega_0$ for three values of ζ. Fig. 16 shows graphs for $\varepsilon^{-\zeta\omega_0 t}\sin\omega_0 t$. Since eqn. (3) is linear we may add these functions to obtain more complicated graphs, taking care that the physical conditions are met.

These results suggest a means of approximating analytically to a frequency response $\mathrm{F}(j\omega)$ given by experiment. We may plot

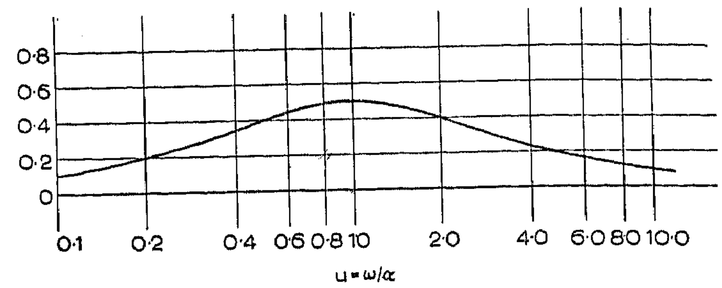


Fig. 14.—Imaginary part of the frequency function corresponding to $f(t) = \varepsilon^{-\alpha t}$.

The curve is plotted against log u, where $u = \omega/\alpha$.

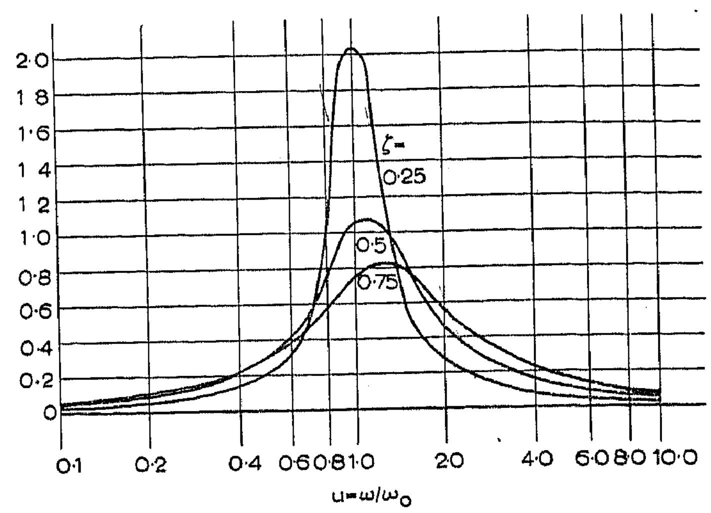


Fig. 15.—Imaginary part of the frequency function corresponding to $f(t) = \varepsilon^{-\zeta\omega_0 t} \cos \omega_0 t$.

The curves are plotted against log u, where $u = \omega/\omega_0$, for $\zeta = 0{\cdot}25$, $0{\cdot}5$ and $0{\cdot}75$.

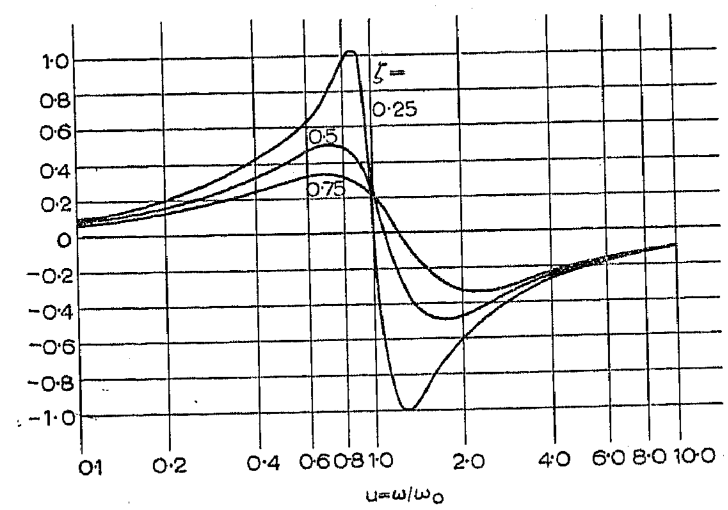


Fig. 16.—Imaginary part of the frequency function corresponding to $f(t) = \varepsilon^{-\zeta\omega_0 t} \sin \omega_0 t$.

The curves are plotted against log u, where $u = \omega/\omega_0$, for $\zeta = 0{\cdot}25$ $0{\cdot}5$ and $0{\cdot}75$.

$\mathscr{I}\mathrm{F}(j\omega)$ and then subtract from it an appropriate multiple of the graph which seems to fit it most closely. This gives one mode in the approximation. Repeating the procedure we may successively reduce the remainder, and at any stage by means of the cursor we may find the transient response due to this remainder. As soon as this response is negligible for the conditions of the given problem, we may regard the approximate analytical representation as complete. The advantage of this procedure is that we can work in the frequency domain where the frequency components of the response are easily identified, and where a change of time scale involves only translation along the scale of log ω. At the same time we have an immediate check on the significance of the remainder in the time domain.

(6) DIRECT LAPLACE TRANSFORMATION

If we are given $f(t)$ and wish to find $\mathscr{I}\mathrm{F}(j\omega)$, we may make use of the cursor already described. We plot $At[f(t) - f(\infty)]$ against log t to the same scales as were used in Section 2 for $\mathscr{I}\mathrm{F}(j\omega)$ and log ω respectively, A being a convenient constant. We then make use of the cursor, setting ω on the scale for t. The result, when multiplied by $\pi\omega/200A$, gives $\mathscr{I}\mathrm{F}(j\omega)$, as shown in Section 12. It is possible to obtain $\mathscr{R}\mathrm{F}(j\omega)$ in a similar way by the use of a cursor derived from sin ωt, and so to determine $\mathrm{F}(j\omega)$.

(7) FURTHER APPLICATIONS

The approximate method given in Section 2 has been presented in its application to automatic control systems with a step-function input and with particular initial conditions. It has, however, more general applications, and some of these are suggested below.

(7.1) Response to other Inputs

If the system is subjected to an input having Laplace transformation $\theta_i(p)$ we may find the response by using $\mathscr{I}[j\omega\theta_i(j\omega)\mathrm{F}(j\omega)]$ in place of $\mathscr{I}[\mathrm{F}(j\omega)]$ in eqn. (3).

(7.2) Other Initial Conditions

It is assumed in eqn. (1) that the system is initially at rest, i.e. in the condition which it reaches after a long time when subject to no disturbances. For other initial conditions the function $\mathrm{F}(p)/p$ is modified according to well-known rules,[1] and the method given in Section 2 can be applied to the modified function.

(7.3) Other Linear Systems

Eqn. (1) arises not only for servo systems, but for any system governed by linear differential equations, such as electrical networks, vibrating mechanical systems, etc. The method given in Section 2 can be applied in these cases.

(7.3.1) Impulse Testing.

A particular example of an electrical system to which the method might be applied is the impulse generator. The problem here is to calculate the voltage transient appearing across the apparatus under test from the initial conditions and the circuit constants. The transient can be expressed in a form similar to that of eqn. (1), and evaluated as in Section 2.

(7.4) Fourier Cosine Transforms

In eqn. (3), $f(t) - f(\infty)$ is $\sqrt{(\pi/2)}$ times the Fourier cosine transform of $\mathscr{I}\mathrm{F}(j\omega)/\omega$. The method given in Section 2 evaluates the integral in eqn. (3), and can be used to obtain such transforms in general.

(7.4.1) Auto-Correlation Functions.

An example of a Fourier cosine transform arises in finding the quadratic spectrum $\phi(\omega)$ of a function from its auto-correlation function $R(\tau)$. It is well known[11,12] that

$$\phi(\omega) = \frac{2}{\pi}\int_0^\infty R(\tau)\cos\omega\tau\, d\tau \quad . \quad . \quad . \quad . \quad (12)$$

$$= \frac{2}{\pi}\int_{\tau=0}^\infty \tau R(\tau)\cos\omega\tau\, d(\log\tau) \quad . \quad . \quad (13)$$

and

$$R(\tau) = \int_0^\infty \phi(\omega)\cos\omega\tau\, d\omega \quad . \quad . \quad . \quad . \quad (14)$$

$$= \int_{\omega=0}^\infty \omega\phi(\omega)\cos\omega\,\tau d(\log\omega) \quad . \quad . \quad (15)$$

A comparison of eqns. (13) and (15) with eqn. (3) shows how these expressions may be evaluated. In all important cases $R(\infty) = 0$.

(8) LIMITATIONS OF THE METHOD

It is evident from Fig. 9 that the method will fail if there exists a very lightly damped mode in the response. In these circumstances there are simple alternative methods[13,14] for finding the lightly-damped mode. If the Laplace transform corresponding to this mode is subtracted from $\theta_i(p)\mathrm{F}(p)$, the remaining part of the response can be evaluated as in Section 2.

Similar remarks apply when the input function is harmonic, in which case $\theta_i(p)\mathrm{F}(p)$ has poles on the imaginary axis other than at $p = 0$.

(9) CONCLUSIONS

By means of a simple procedure it is possible to evaluate the direct and inverse Laplace-transform integrals [eqns. (25) and (16)] with an accuracy sufficient for many engineering purposes. It is hoped that the method for doing this, which is explained in Section 2, will be useful in the design of control systems, and for some other purposes.

(10) ACKNOWLEDGMENT

The author wishes to express his thanks to the directors of Costain–John Brown, Ltd., for permission to publish this paper.

(11) REFERENCES

(1) CARSLAW, H. S., and JAEGER, J. C.: "Operational Methods in Applied Mathematics" (Oxford University Press, 1949).
(2) WILLIAMS, F. C., and RITSON, F. J. U.: "Electronic Servo Simulators," *Journal I.E.E.*, 1947, **94**, Part IIA, p. 112.
(3) WARD, E. E.: "An Analogue Computer for Use in the Design of Servo Systems," *Proceedings I.E.E.*, Paper No. 1201 M, November, 1951 (**99**, Part II, p. 521).
(4) EVANS, W. R.: "Control System Synthesis by the Root Locus Method," *Transactions of the American I.E.E.*, 1950, **69**, p. 66.
(5) WESTCOTT, J. R.: "The Minimum-Moment-of-Error-Squared: a New Performance Criterion for Servo Mechanisms," *Proceedings I.E.E.*, Paper No. 1644 M, March, 1954 (**101**, Part II, p. 471).
(6) BEDFORD, A. V., and FREDENDALL, G. L.: "Analysis, Synthesis, and Evaluation of the Transient Response of Television Apparatus," *Proceedings of the Institute of Radio Engineers*, 1942, **30**, p. 440.
(7) BROWN, G. S., and CAMPBELL, D. P.: "Principles of Servomechanisms" (J. Wiley, New York, 1948), p. 332. (The method is due to G. F. Floyd.)
(8) HAMOS, L. VON., JANSSON, B., and PERSSON, TH.: "Simplified Methods for the Evaluation of Transients in Linear Systems," *Acta Polytechnica, Physics and Applied Mathematics Series*, 1952, **2**, No. 112.
(9) WAIT, J. R.: "An Approximate Method of obtaining the Transient Response from the Frequency Response," *Canadian Journal of Technology*, 1953, **31**, p. 127.
(10) LEONHARD, A.: "Determination of Transient Response from Frequency Response," *Transactions A.S.M.E. Frequency-response Symposium*, 1954, **76**, p. 1215.
(11) JAMES, H. M., NICHOLS, N. B., and PHILLIPS, R. S.: "Theory of Servomechanisms" (McGraw-Hill, New York, 1947), p. 283.
(12) TAYLOR, G. I.: "The Spectrum of Turbulence," *Proceedings of the Royal Society*, A, 1938, **164**, p. 476.
(13) MACMILLAN, R. H.: "An Introduction to the Theory of Control in Mechanical Engineering" (Cambridge University Press, 1951), p. 132.
(14) WEST, J. C., and POTTS, J.: "A Simple Connection between Closed-Loop Transient Response and Open-Loop Frequency Response," *Proceedings I.E.E.*, Paper No. 1378 M, August, 1952 (**100**, Part II, p. 201).
(15) COPSON, E. T.: "Theory of Functions of a Complex Variable" (Oxford University Press, 1944), p. 135.
(16) COSTAIN-JOHN BROWN, LTD.: British Patent Application No. 2217, 1955.

(12) APPENDIX

If the transfer function of a linear system is $\mathrm{F}(p)$, having no poles in the right-hand half-plane or on the imaginary axis, the time response of the system to a unit step function is[1]

$$\mathrm{f}(t) = \frac{1}{2\pi j}\int_{-j\omega}^{j\omega} \frac{\mathrm{F}(p)\varepsilon^{pt}}{p}dp \quad . \quad . \quad . \quad . \quad (16)$$

The path of integration is the imaginary axis, with an indentation into the right-hand half-plane, if necessary, to avoid a pole at the origin. As the radius of this indentation is reduced the integral around it tends[15] to one-half of the residue at the pole. Moreover, this residue is, under the assumed conditions, the limiting value of $\mathrm{f}(t)$ as t tends to infinity. Thus we obtain

$$\mathrm{f}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\mathrm{F}(j\omega)\varepsilon^{j\omega t}}{j\omega}d\omega + \tfrac{1}{2}\mathrm{f}(\infty) \quad . \quad . \quad . \quad (17)$$

where the integral has its Cauchy principal value.

Owing to the even symmetry of the real part of the integrand in eqn. (17), and the odd symmetry of the imaginary part, it follows that

$$\mathrm{f}(t) = \frac{1}{\pi}\int_{0}^{\infty} \mathscr{R}\left[\frac{\mathrm{F}(j\omega)\varepsilon^{j\omega t}}{j\omega}\right]d\omega + \tfrac{1}{2}\mathrm{f}(\infty) \quad . \quad . \quad (18)$$

Moreover, it follows from this derivation that the integral in eqn. (18) converges at the lower limit. Now

$$\mathscr{R}\left[\frac{\mathrm{F}(j\omega)\varepsilon^{j\omega t}}{j\omega}\right] = \mathscr{R}\left\{\left[\frac{\mathscr{R}\mathrm{F}(j\omega) + j\mathscr{I}\mathrm{F}(j\omega)}{j\omega}\right](\cos\omega t + j\sin\omega t)\right\}$$
$$= \frac{1}{\omega}[\mathscr{I}\mathrm{F}(j\omega)\cos\omega t + \mathscr{R}\mathrm{F}(j\omega)\sin\omega t] \quad . \quad (19)$$

so that

$$\mathrm{f}(t) = \frac{1}{\pi}\int_{0}^{\infty}[\mathscr{I}\mathrm{F}(j\omega)\cos\omega t + \mathscr{R}\mathrm{F}(j\omega)\sin\omega t]\frac{d\omega}{\omega} + \tfrac{1}{2}\mathrm{f}(\infty) \quad . \quad . \quad . \quad . \quad (20)$$

When t is negative, $\mathrm{f}(t)$ is identically zero. It follows from the odd and even symmetry of $\sin\omega t$ and $\cos\omega t$ respectively that for positive values of t

$$\frac{1}{\pi}\int_{0}^{\infty}\mathscr{I}\mathrm{F}(j\omega)\cos\omega t\frac{d\omega}{\omega} + \frac{1}{2}\mathrm{f}(\infty) = \frac{1}{\pi}\int_{0}^{\infty}\mathscr{R}\mathrm{F}(j\omega)\sin\omega t\frac{d\omega}{\omega} \quad . \quad . \quad . \quad . \quad (21)$$

whence

$$\left.\begin{aligned}\mathrm{f}(t) &= \frac{2}{\pi}\int_{\omega=0}^{\infty}\mathscr{R}\mathrm{F}(j\omega)\sin\omega t\, d(\log\omega)\\ \mathrm{f}(t) - \mathrm{f}(\infty) &= \frac{2}{\pi}\int_{\omega=0}^{\infty}\mathscr{I}\mathrm{F}(j\omega)\cos\omega t\, d(\log\omega)\end{aligned}\right\} \quad . \quad (22)$$

When $\mathrm{F}(j\omega)$ is the closed-loop harmonic response of a servo system with unity feedback it is given explicitly by

$$\mathrm{F}(j\omega) = G(\omega)/[1 + G(\omega)] \quad . \quad . \quad . \quad . \quad (23)$$

where $G(\omega)$ is the open-loop harmonic response. If $\mathrm{F}(j\omega)$ corresponds to the deviation $\theta_i - \theta_0$, it is given by

$$\mathrm{F}(j\omega) = 1/[1 + G(\omega)] \quad . \quad . \quad . \quad . \quad (24)$$

In the second case there will be no pole of $\mathrm{F}(j\omega)/\omega$ at the origin if $G(\omega)$ tends to infinity as ω tends to zero.

To obtain $\mathrm{F}(j\omega)$ from $\mathrm{f}(t)$ we observe that

$$\frac{\mathrm{F}(p)}{p} = \int_0^\infty \mathrm{f}(t)\varepsilon^{-pt}dt \quad . \quad . \quad . \quad . \quad . \quad (25)$$

and

$$\frac{\mathrm{f}(\infty)}{p} = \int_0^\infty \mathrm{f}(\infty)\varepsilon^{-pt}dt \quad . \quad . \quad . \quad . \quad (26)$$

Thus

$$\frac{\mathrm{F}(p)}{p} - \frac{\mathrm{f}(\infty)}{p} = \int_0^\infty [\mathrm{f}(t) - \mathrm{f}(\infty)]\varepsilon^{-pt}dt \quad . \quad . \quad (27)$$

Since $\mathrm{f}(t) - \mathrm{f}(\infty)$ tends to zero as t tends to infinity, the integral in eqn. (27) remains convergent if we put $p = j\omega$. Thus

$$\mathrm{F}(j\omega) = \mathrm{f}(\infty) + j\omega\int_0^\infty [\mathrm{f}(t) - \mathrm{f}(\infty)](\cos\omega t - j\sin\omega t)dt \quad (28)$$

and

$$\mathscr{I}\mathrm{F}(j\omega) = \omega\int_0^\infty [\mathrm{f}(t) - \mathrm{f}(\infty)]\cos\omega t dt$$

$$= \frac{\pi\omega}{200}\left\{\frac{200}{\pi}\int_{t=0}^\infty t[\mathrm{f}(t) - \mathrm{f}(\infty)]\cos\omega t\, d(\log t)\right\} \quad (29)$$

The quantity within curly brackets in eqn. (29) is the result obtained by plotting $t[\mathrm{f}(t) - \mathrm{f}(\infty)]$ against $\log t$ and using the cursor as explained in Section 2. We have similarly

$$\mathscr{R}\mathrm{F}(j\omega) = \mathrm{f}(\infty) + \omega\int_0^\infty [\mathrm{f}(t) - \mathrm{f}(\infty)]\sin\omega t dt$$

$$= \mathrm{f}(\infty) + \frac{\pi\omega}{200}\left\{\frac{200}{\pi}\int_{t=0}^\infty t[\mathrm{f}(t) - \mathrm{f}(\infty)]\sin\omega t\, d(\log t)\right\}$$

$$. \quad . \quad . \quad . \quad . \quad (30)$$

DISCUSSION ON "STANDARD WAVEGUIDES AND COUPLINGS FOR MICROWAVE EQUIPMENT"*

Mr. M. S. Seaman (*communicated*): Table 1 shows the attenuation of the various sizes of standard waveguides, but the author does not specify the material with which these figures are obtained, nor does he say whether they are theoretical or practical values.

For the WG 20 size, the figure of 10·9 dB/100 ft which he quotes appears to be the attenuation which should theoretically be obtained in a copper guide. However, precision-drawn copper guides have been found to give an attenuation considerably in excess of this, typical figures being about 13 dB/100 ft—a value which has also been obtained in America by Maxwell.† This increase in the attenuation is due to the surface roughness of the internal walls of the guide, as has been shown by Benson (see Reference 9). For a brass guide of this size a typical attenuation obtained with precision-drawn guides is about 21 dB/100 ft, compared with a theoretical value of 19·9 dB/100 ft.

* HARVEY, A. F.: Paper No. 1807 R, July, 1955 (see Vol. **102** B, p. 493).
† MAXWELL, E.: "Conductivity of Metallic Surfaces at Microwave Frequencies," *Journal of Applied Physics*, 1947, 18, p. 629.

Dr. A. F. Harvey (*in reply*): Only brief technical data are given in Table 1, and References 16, 17 and 21 should be consulted for further information. The attenuation figures are for copper or silver material with a resistivity of $1\cdot62 \times 10^{-6}$ ohm-cm and are theoretical values for a frequency approximately 1·5 times the cut-off value. For other frequencies the figures quoted should be multiplied by the factor

$$0\cdot421[(f/f_c)^2 + 1]/(f/f_c)^{\frac{1}{2}}[(f/f_c)^2 - 1]^{\frac{1}{2}}$$

where f and f_c are respectively the actual and cut-off frequencies. For other materials the figures quoted should be multiplied by 1·06 (silver with 7½% copper), 1·30 (aluminium), 1·55 (brass with 90% copper) and 2·00 (brass with 70% copper). The increase in attenuation caused by surface roughness is held within about 20% by specifying a maximum r.m.s. value of one-half the electrical skin depth at the mean operating frequency in the given material. If these factors are applied, the agreement with the practical values quoted by Mr. Seaman is seen to be very good.

Paper No. 1811 R
Mar. 1955

621.317.34 : 621.314.7

MEASUREMENTS OF JUNCTION-TRANSISTOR NOISE IN THE FREQUENCY RANGE 7–50 KC/S

By W. L. STEPHENSON, B.Sc.

(*The paper was first received* 13*th September, and in revised form* 21*st December*, 1954. *It was published in March*, 1955, *and was read before a* JOINT MEETING *of the* RADIO *and* MEASUREMENTS SECTIONS 11*th May*, 1955.)

SUMMARY

The paper describes an investigation to determine the variation of junction transistor noise with operating conditions (frequency, temperature, voltage and current) in the frequency range 7–50 kc/s.

A detailed investigation involves a statistical analysis of a very large number of measurements, but since in this case it was required to determine only approximate laws, a small number of measurements were taken with the assumption of a large possible error.

The noise voltages were measured using the superheterodyne principle to give constant bandwidth over the frequency range.

The analysis of the results is made using the principle of two equivalent noise generators in the input circuit: theoretical considerations link these with the minimum noise factor and optimum source resistance, which together determine the noise characteristics of the transistor.

LIST OF PRINCIPAL SYMBOLS

A = Transistor voltage gain.
f_{b2} = Bandwidth, c/s.
F = Noise factor.
F_{min} = Minimum noise factor when R_s is varied.
G_V = Equivalent noise-voltage generator.
G_I = Equivalent noise-current generator.
I_g = Current output of G_I.
k = Boltzmann's constant ($1 \cdot 38 \times 10^{-23}$ joule/°K).
$K = \sqrt{(4kTf_{b2})}$.
R_{in} = Transistor input resistance.
R_s = Resistance of the signal source from which the transistor is driven.
R_{opt} = Value of R_s at which $F = F_{min}$.
T = Temperature, °K.
V_g = Voltage output of G_V.
V_s = Source voltage applied to the transistor input terminals.

(1) INTRODUCTION

Figures from many sources have been published showing that transistor noise falls at 3 dB per octave with increasing frequency, but this law can hold only over a limited frequency range and it is not generally understood at what frequency the law is no longer valid, or what law supersedes it beyond that point.

This investigation was undertaken, therefore, to determine the noise characteristics of junction transistors at frequencies above the audio-frequency range. The transistors measured were of the *p-n-p* fused-junction type with a cut-off frequency of approximately 400 kc/s.

Noise is a random phenomenon, and its level can be accurately determined only from its statistical properties. Such a determination would involve the measurement of a large number of transistors, each measurement being effected many times. In this case, however, we do not want such a detailed examination but merely to determine approximate laws and the orders of magnitude involved. Thus it is sufficient to take measurements on a small number of transistors (about ten) and to assume a large possible random error; the results will give an idea of the magnitude of variation of noise with operating conditions (frequency, temperature, current, and voltage), and approximate laws can be established.

The consideration of the exact source of noise in the transistor is of purely academic interest to the circuit engineer since, in general, he uses the transistor as a whole and is satisfied if he can determine from published data and standard formulae the noise figure of the transistor 4-pole network at the required operating conditions.

(2) MEASUREMENT OF NOISE VOLTAGE

(2.1) Method of Measurement

The method of measurement of noise voltage is a well-known technique; the noise is amplified and passed through a filter of known pass-band, the output being measured with a voltmeter.

The obvious method of measuring noise voltages over a range of frequencies is then to have a series of bandpass filters of known bandwidths centred at different frequencies over the range to be measured. This method, however, is clumsy and not amenable to simple calculations, since the bandwidth will vary with the frequency at which the measurements are taken.

A more elegant method is to use the superheterodyne principle with a narrow-band filter in the i.f. amplifier. The problems presented by this method are basically those of the superheterodyne receiver, namely the rejection of i.f. break-through and second-channel interference. We can therefore draw up a specification of the measuring equipment required.

(2.1.1) Filter Bandwidth.

Since the noise voltages vary with frequency, it is essential that the bandwidth be small compared with the frequency at which the measurements are taken. The smaller the bandwidth the more nearly will the measurement be that at a "spot" frequency, but the smaller will be the output voltage to be measured. In practice a 2-stage tuned-ladder filter can give a bandwidth of approximately 0·5% of the mid-band frequency. However, such a narrow band is not necessary, and a bandwidth of 5% of the mid-band frequency is sufficient. This then limits the minimum frequency covered by the equipment to one-tenth of the intermediate frequency.

(2.1.2) The Intermediate Frequency.

From the nature of the frequency band (7–50 kc/s) to be covered it is essential that the intermediate frequency be above the maximum frequency to be measured, and therefore that the oscillator frequency be above the intermediate frequency.

Since the pre-amplifier cannot easily be made very selective at the low frequencies to be measured, it is essential that a tuned filter be incorporated to reject all incoming signals at the intermediate frequency.

(2.1.3) The Voltmeter.

The i.f. output should be measured with a voltmeter which reads the mean or the r.m.s. voltages, since noise contains dis-

Mr. Stephenson is at the Mullard Research Laboratories.

proportionately large peaks and a peak-reading voltmeter will therefore give an inaccurate reading if calibrated with a sinusoidal waveform.

A further difficulty that occurs is that, since the noise level of a transistor falls with increasing frequency, a wide-band pre-amplifier which is sufficiently sensitive to measure the very low voltages that occur at the high frequencies will be overloaded by the higher voltages at the low-frequency end of the band which occur at the same time. This overloading will produce harmonics, which add to the high-frequency noise being measured, and thereby produce inaccuracies. The effect can be eliminated by selective tuning, so that the pre-amplifier overall response curve gives a rejection of at least 3 dB per octave from the selected frequency.

Fig. 1 is a block diagram of the system used, and some stages of it will now be described in greater detail.

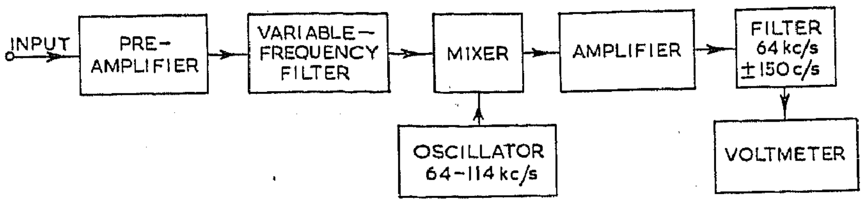


Fig. 1.—Diagram of noise-measuring equipment.

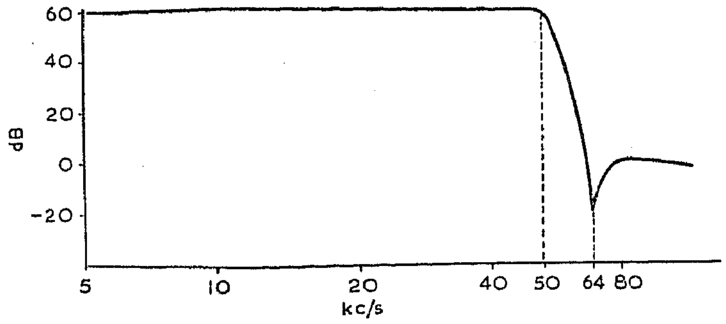


Fig. 2.—Pre-amplifier gain characteristic.

The pre-amplifier incorporates a filter (Fig. 2) having a flat response to 50 kc/s and giving more than 60 dB rejection at and above 64 kc/s (the intermediate frequency) to prevent i.f. and harmonic break-through.

The mixer used (Fig. 3) is a double balanced bridge modu-

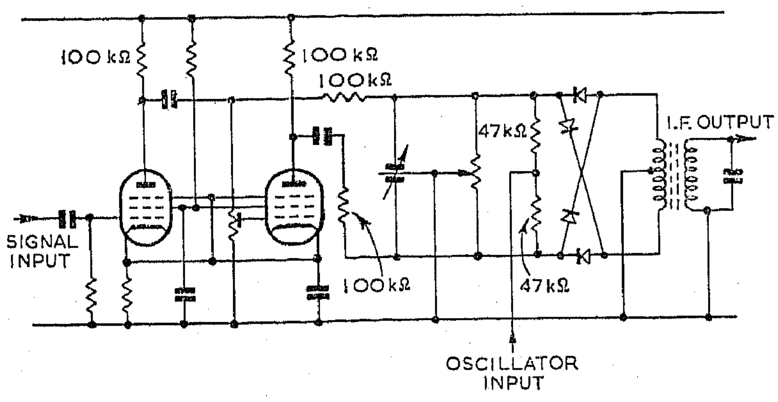


Fig. 3.—Circuit of the mixer stage.

lator using germanium diodes and fed from a valve phase-splitter. This mixer circuit is very critical on balance, a small change in grid resistance at the input to the phase splitter causing a change in balance conditions owing to Miller effect in the phase-splitter pentodes. It is possible that a different type of phase splitter may overcome this effect, but the circuit used was considered satisfactory.

In fact it is not essential that the mixer be of the balanced type provided that the selectivity of the i.f. amplifier is sufficient to reject the oscillator-frequency output of the mixer.

The filter is a 2-stage π-section tuned-ladder filter with a bandwidth of 300 c/s and a mid-band frequency of 64 kc/s.

The output of the i.f. amplifier is measured with a double-triode valve voltmeter (Fig. 4) having a short time-constant in

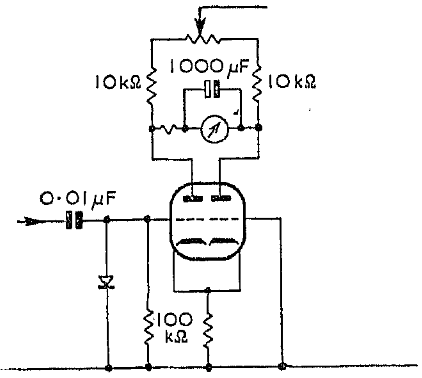


Fig. 4.—Circuit of the valve voltmeter.

the rectifier circuit and a long time-constant in the meter circuit, thereby reading the average value of mean voltage over a long period.

In this equipment the usual precautions with regard to screening against unwanted noise were observed, and the basic noise level was equivalent to 0·2 microvolt at the input, making it possible to read levels of the order of 1 microvolt.

(3) MEASUREMENT OF NOISE FACTOR

The use of the noise factor is generally considered the simplest method of expressing the noise characteristics of an amplifier, since it is the ratio of the signal/noise ratios at the input and output.

The noise factor, F, of an amplifier is defined as the ratio of the actual noise power in the load to the theoretical noise power in the load due to Johnson noise in the source resistance, assuming the amplifier to be noiseless.

This can be expressed as

$$F = (V_n/V_sA)^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

Now a transistor has a finite input resistance, so that the input noise voltage is not the open-circuit noise of the source resistance R_s. However, we can consider Johnson noise to be due to a voltage generator of amplitude $K\sqrt{(R_s)}$ where $K = \sqrt{(4kTf_{b2})}$, with an internal resistance R_s (Fig. 5).

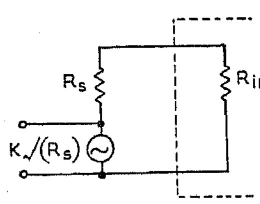


Fig. 5.—Equivalent input circuit for the measurement of noise factor.

The method of measurement of noise factor is then to measure the voltage gain of the amplifier from the generator terminals; if this is A' and the noise voltage at the output terminals is V_n, then

$$F = [V_n/K\sqrt{(R_s)}A']^2 \quad . \quad . \quad . \quad . \quad . \quad (2)$$

(4) EQUIVALENT NOISE GENERATORS

(4.1) Limitations of the Noise-Factor Principle

From eqn. (1) we can see that knowing the noise factor we can calculate the actual noise output:

$$V_n = V_sA\sqrt{F} \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

Unfortunately this equation is true only if R_s approximates to the value used in the measurement of F, since the noise generated by the transistor is not necessarily proportional to the noise generated in the source resistance.

Now the transistor is a linear current-amplifier, and in the interests of linearity we frequently require the transistor to be driven from a constant-current (high-impedance) source. In such a case, the knowledge of the noise factor with a source resistance of 500 ohms is of little help in the determination of the amplifier noise characteristics, so that another method of describing noise is required, which is independent of the source resistance.

(4.2) Noise in 4-Pole Networks

By a theorem of Peterson* we can consider any noisy 4-pole network to be equivalent to a noiseless 4-pole network having the same parameters together with two fictitious noise generators, one being in the input circuit and the other in the output circuit. A corollary of this, due to Becking, is that if the 4-pole network has a transmission characteristic the generators may both be situated in the input circuit.

We can therefore consider our transistor to have an equivalent input circuit as shown in Fig. 6.

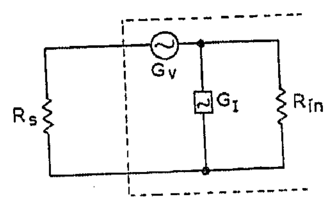


Fig. 6.—Input circuit with two equivalent noise generators.
Gv generates a constant noise voltage, V_g.
GI generates a constant noise current, I_g.
There may be some correlation between V_g and I_g.

We can measure the values of V_g and I_g by measuring the input resistance, R_{in}, and the voltage gain of the transistor and by measuring the output noise voltage for $R_s = 0$ and $R_s = \infty$. If we then give R_s finite values we can determine to what degree V_g and I_g are correlated, and can determine a method of calculating the noise characteristics of the transistor for any value of V_g and I_g.

In fact, measurements show that there is little, if any, correlation between these two noise generators, and that knowing V_g, I_g, A and R_{in} we can predict the noise output for any value of R_s.

(4.3) Theoretical Considerations

If we consider again the circuit shown in Fig. 6 the total noise voltage is

$$\frac{V_g + K\sqrt{R_s} + I_g R_s}{R_{in} + R_s} R_{in}$$

The input noise voltage due to R_s is

$$\frac{K\sqrt{R_s}}{R_{in} + R_s} R_{in}$$

So that the noise factor is

$$\left(\frac{V_g + K\sqrt{R_s} + I_g R_s}{K\sqrt{R_s}}\right)^2$$

$$F = \left[1 + \frac{V_g}{K\sqrt{R_s}} + \frac{I_g\sqrt{R_s}}{K}\right]^2$$

If we let

$$R_s = n^2\frac{V_g}{I_g} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

then

$$F = \left[1 + \frac{\sqrt{(V_g I_g)}}{K}\left(n + \frac{1}{n}\right)\right]^2 \quad . \quad . \quad . \quad . \quad (5)$$

which is a minimum when $n = 1$; so that the optimum source resistance is V_g/I_g and the minimum noise factor is $[1 + 2\sqrt{(V_g I_g)}/K]^2$. By means of eqns. (4) and (5) and knowing R_{opt} and F_{min} we can determine the noise factor for any value of R_s.

The variation of F with R_s/R_{opt} for various values of F_{min} is shown in Fig. 7. This is seen to agree very closely with published experimental results.

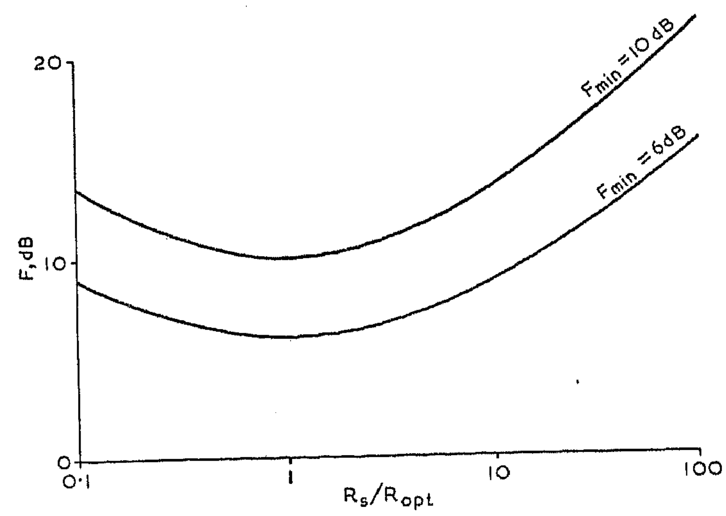


Fig. 7.—Theoretical variation of noise factor with source resistance.

(5) RESULTS

Figs. 8–11 show the variations of F_{min}, R_{opt}, R_{in}, V_g and I_g for variation of frequency, temperature, collector current and collector voltage on earthed-emitter transistors.

In these, V_g and I_g are referred to 1 c/s bandwidth, assuming that both are proportional to the square root of the bandwidth when this is small compared with the frequency. In the computing of the actual noise figure it is simpler to use the para-

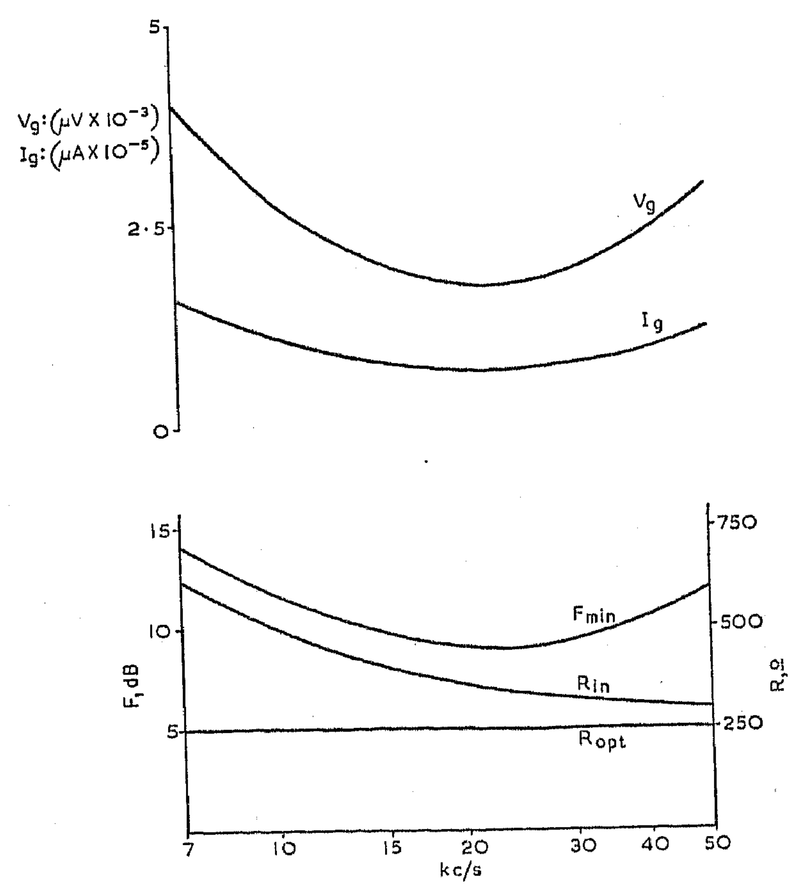


Fig. 8.—Variation of noise parameters with frequency.
$V_c = 2$ V, $I_c = 2$ mA, $T = 22°$ C.

* MONTGOMERY, H. C.: *Proceedings of the Institute of Radio Engineers*, 1952, 40, p. 1461.

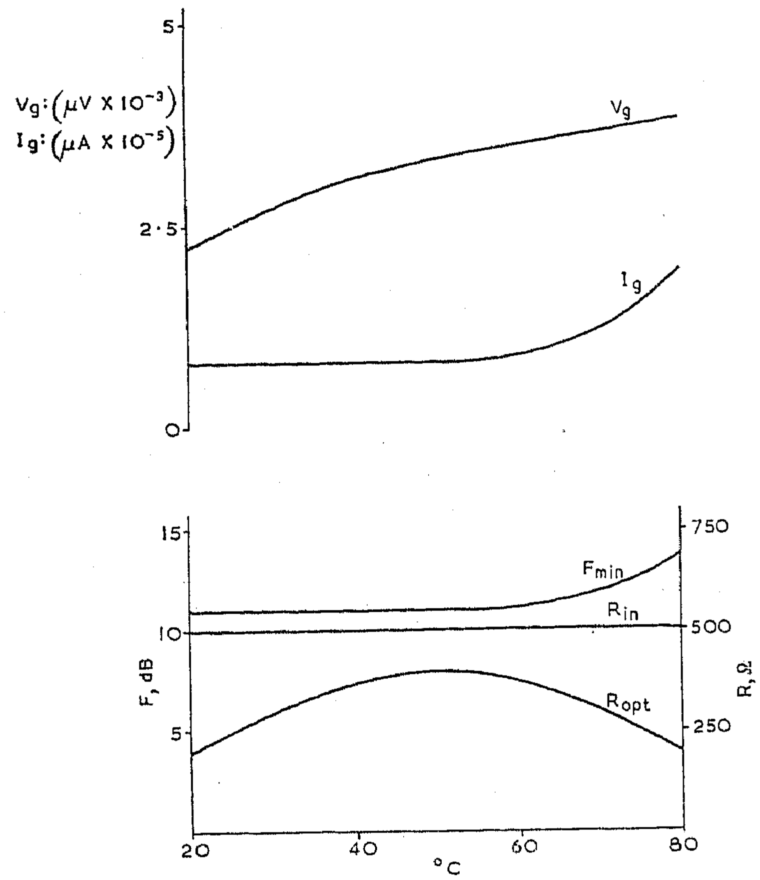


Fig. 9.—Variation of noise parameters with temperature.
V_c=2 V, I_c=2 mA, f=10 kc/s.

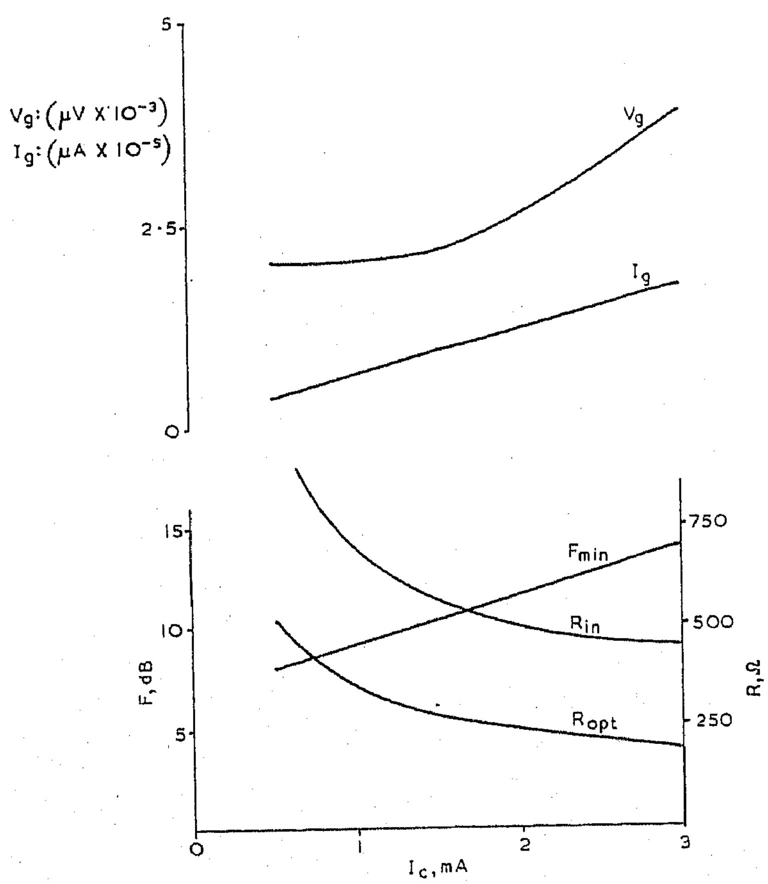


Fig. 10.—Variation of noise parameters with collector current.
V_c=2 V, f=10 kc/s, T=22° C.

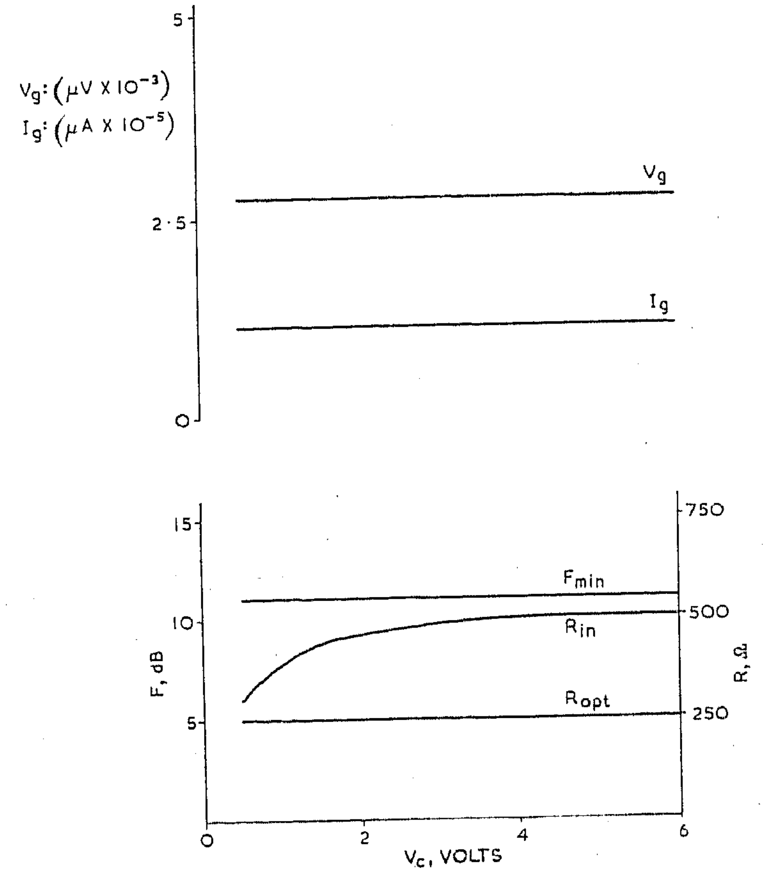


Fig. 11.—Variation of noise parameters with collector voltage.
I_c=2 mA, f=10 kc/s, T=22° C.

meters F_{min} and R_{opt} than V_g and I_g, since the former are independent of bandwidth if this is small compared with the frequency; moreover, the value of the transistor input resistance is not required in the calculation.

The method of obtaining the actual noise figure from the graphs is best shown by means of an example.

To find the actual noise figure under the conditions

$$I_c = 1 \cdot 5\,\text{mA},\ R_s = 2\,500\ \text{ohms},\ f = 10\,\text{kc/s}$$

$$\text{From Fig. 10: } F_{min} = 10\,\text{dB},\ R_{opt} = 250\ \text{ohms},$$

so that $R_s/R_{opt} = 10$.

Referring to Fig. 7, we find that under these conditions $F = 15$ dB, which is the true noise figure.

The graphs of variation with temperature and frequency can be used in a similar manner.

(6) CONCLUSIONS

The most interesting conclusions to be drawn from this investigation are:

(*a*) The noise factor reaches a minimum at about 25 kc/s as the frequency is increased, and rises beyond this point.

(*b*) The optimum source resistance is independent of frequency.

(*c*) The optimum source resistance is not directly related to the input resistance.

(*d*) The noise is completely independent of collector voltage.

The last of the above conclusions is particularly interesting, since it directly contradicts published results of many other experiments. The reason is almost certainly that previous figures refer to the early transistors which were not hermetically sealed.

[The discussion on the above paper will be found on page 786.]

Paper No. 1860 R
Apr. 1955

621.383 : 669.783

A GERMANIUM DIFFUSED-JUNCTION PHOTO-ELECTRIC CELL

By J. M. WADDELL, M.A., Graduate, S. E. MAYER, Ph.D., and S. KAYE, B.Sc., Student.

(*The paper was first received* 4*th March and in revised form* 18*th March*, 1955. *It was published in April*, 1955, *and was read before a* JOINT MEETING *of the* RADIO *and* MEASUREMENTS SECTIONS 11*th May*, 1955.)

SUMMARY

The paper describes a new type of germanium-junction photo-electric cell, in which the junction is manufactured by impurity diffusion. This type of cell differs from those made by other methods in having a lower noise level and a slightly extended long-wave response. The possible causes of this latter anomaly are discussed, and some applications are suggested which utilize both effects. Two practical photocells are described which differ mainly in physical dimensions. Both are hermetically sealed and have high sensitivity and large output for their size. It is shown that the maximum reverse voltage that can be applied to a germanium junction device is a function of the cooling arrangements provided. The special case of the photocell is considered, and it is shown that the criterion for thermal stability is temperature rise rather than absolute temperature. For a germanium photocell in the dark the maximum temperature rise is approximately 11°C.

LIST OF SYMBOLS

T = Absolute temperature, °K.
ΔT = Temperature rise, °K.
k = Boltzmann's constant.
W_i = Intrinsic energy gap, eV.
P_l = Power loss in a junction.
P_d = Power dissipated from a junction.
f = Frequency, c/s.
C = Capacitance of a junction.
V = Voltage applied to a junction.
$\frac{dN}{df}$ = Noise power per unit bandwidth.
I_s = Reverse saturation current.
K, m, n = Various constants.
P_1 = Temperature-dependent losses.
P_2 = Temperature-independent losses.

(1) INTRODUCTION

When a semi-conductor such as germanium is illuminated by light of suitable wavelength, pairs of oppositely charged free current carriers, holes and electrons, are generated. Thus, without the semi-conductor acquiring any net charge, the total number of free carriers present is increased and the resistivity is lowered. This effect was first observed[1] in selenium in 1873. The resistivity of germanium in the dark is too low to allow its practical use as a photo-conductive cell, but a number of other photo-conductive materials having high resistivities in the dark have been used for photocells in recent years, especially since some of them respond to light of relatively long wavelength.[2]

If the light falls on the semi-conductor in the immediate neighbourhood of a rectifying barrier, a voltage is built up across the barrier; if a load is connected across the terminals of the photocell, power can be obtained from the light beam itself. Photocells of the copper-oxide and selenium barrier-layer types[3] have used this effect for the measurement of illumination, photographic exposure, etc., for many years, while recently a silicon *p–n*-junction cell has been described for the direct conversion of solar to electrical power.[4]

In nearly all germanium photocells, rectifying barriers are used in a somewhat different manner; the cell is operated in the photo-conductive mode, with the barrier biased in the reverse direction so as to overcome the low resistivity of the germanium and make the dark current comparable with, or less than, the light current. Barriers of the point-contact, *p–n* and *n–p–n* types may all be used in this way.[5,6] Of these, the *p–n* junction-type photocell has the lowest dark current and highest working voltage and produces one electron in the external circuit for each photon absorbed immediately adjacent to the barrier;[7] this leads to a high and stable sensitivity. The point-contact type of photocell has a much higher dark current but a somewhat higher if less stable sensitivity; this is due to current multiplication, similar to that at the collector of a point-contact transistor. The *n–p–n* type, or photo-transistor, has a much larger current amplification factor owing to the mechanism known as a "*p–n* hook."[18] This type has an extremely high sensitivity, intermediate dark-current properties, and is somewhat difficult to make reproducibly.

The cells described in the paper are simple *p–n* junction photo-diodes. They have nevertheless very high sensitivity in comparison with the older established types of photocell and they are very small.

(2) MECHANICAL CONSTRUCTION

Owing to the finite lifetime of the minority carriers generated by the light, the sensitivity of a *p–n* junction photocell is a maximum in the immediate neighbourhood of the barrier. To obtain the maximum illumination of the junction, the germanium on one side of the junction (e.g. the *p* side) may be made very thin and the light directed through this. Under these conditions it is difficult to achieve good electrical properties, while, if the layer is made thicker, most of the hole-electron pairs are formed at some distance from the junction and low sensitivity results. In the photocells described, the light is directed on to one edge of the junction; this eliminates the loss due to a thick *p* layer, but the part of the junction remote from the light contributes little to the photo-sensitivity while still adding to the dark current. As a compromise, in the larger of the two photocells described the germanium element is made rectangular, of dimensions 0·060in × 0·035in × 0·020in, the junction being parallel to the largest face. As a result of the method of manufacture of the junctions, the diffusion length of electrons in the *p* layer is appreciably less than that of holes in the *n* layer; the sensitivity of the cell falls off appreciably at distances from the barrier large compared with a diffusion length, and therefore the junction is made asymmetrical, with the *p* layer approximately 0·006in thick, and the *n* layer 0·014in thick.

(2.1) Considerations of Thermal Stability

One of the more important advantages of the *p–n* junction photocell is its ability to produce large signals, either of current

Mr. Waddell is, and Dr. Mayer and Mr. Kaye were formerly, with Standard Telephones and Cables, Ltd.
Mr. Kaye is now with J. Lyons and Co., Ltd.
Dr. Mayer is with the Transitron Electric Corporation, America.

or voltage.[7] In order to produce a large voltage signal, the junction must withstand high reverse direct bias voltages, at least in the dark. Zener breakdown may limit the reverse voltage, but the junction can usually be designed so that the Zener voltage is well outside the working range. The maximum working voltage is then determined by the effect of reverse heating.

Consider any rectifying barrier biased in the reverse direction and in thermal equilibrium with its surroundings. Let its temperature be raised by a small amount, from any cause. The reverse current will increase, and hence the reverse losses; so will the rate of loss of heat from the barrier to the surroundings. If the increase in reverse loss is less than the increase in dissipation, the complete device will remain thermally stable. If the increase in loss is greater than the increase in dissipation, the device will no longer be stable and reverse runaway will take place.[8]

In the case of germanium *p–n* junctions the reverse current is almost independent of the applied voltage at constant temperature, and the variation of the reverse saturation current with temperature of practical junctions agrees quite closely with the theoretical predictions. In the special case of the photocell considered, a continuous reverse bias is applied, and these simplifications lead to a simple and very useful result when the cell is in the dark. The various published expressions[9,10] for the reverse saturation current can all be reduced to the form

$$I_s = KT^n \exp\left(-\frac{W_i}{kT}\right) \quad . \quad . \quad . \quad . \quad . \quad (1)$$

where K and n are constants independent of temperature, and W_i is the intrinsic energy gap. If a voltage V is applied to the junction, the power loss is

$$P_l = VI_s \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

For small temperature rises, the loss of heat from the device is proportional to the temperature rise:

$$P_d = m\Delta T \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

where m is a constant which depends on the cooling arrangement ("fin") used. The junction will be in stable equilibrium if

$$P_d = P_l \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

and

$$\frac{dP_d}{dT} \geqslant \frac{dP_l}{dT} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

When these equations are solved, inserting values appropriate to germanium, it is found that ΔT has a maximum value of approximately 11°C for thermal stability. This is almost independent of the ambient temperature and the values assigned to m and n. It should be noted that this does not imply absolute limits to the junction temperature, the power dissipation or the reverse voltage. These are determined respectively by the ambient temperature, the value of m and the reverse-current characteristic of the junction in question. But it will be seen that it does give a reliable method of predicting the performance of a given junction under given conditions. For example, if the thermal conductance m in the mounting of a germanium element can be doubled, the permissible heat dissipation before runaway occurs can also be doubled. If the reverse current were truly independent of voltage (under isothermal conditions), the maximum reverse voltage could be doubled. Use can be made of this result in the experimental determination of the relative m-values for various mounting arrangements; the power dissipated in the junction at runaway is readily measured for each example, and we do not need to measure the actual junction temperature or the voltage/current characteristics of the various elements.

It should be noted that the particular value of 11°C for ΔT applies only in the case considered, where the heating is due to reverse losses only. Where only part of the losses, say P_1, are temperature dependent, and the remaining losses, say P_2, are relatively independent of temperature, e.g. additional reverse

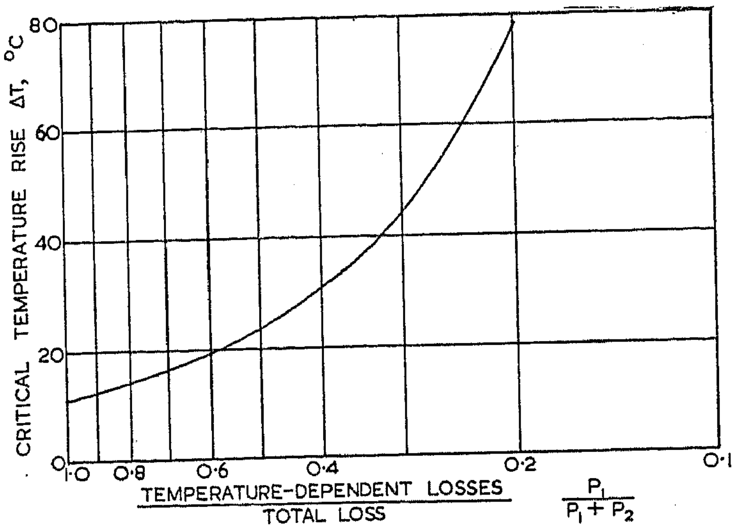


Fig. 1.—Critical temperature rise for a germanium *p–n* junction device as a function of the ratio of temperature-dependent losses to total losses.

current owing to illumination, the maximum temperature before runaway occurs is raised. Fig. 1 shows a graph of the critical temperature rise versus $P_1/(P_1 + P_2)$.

(2.2) Mounting

It is now generally agreed that germanium junction devices should be hermetically sealed if they are to have the long lives of which they are inherently capable, even if they are intended for use in temperate climates. Thus the construction shown in Fig. 2 has been adopted. It is based on a metal-to-glass seal

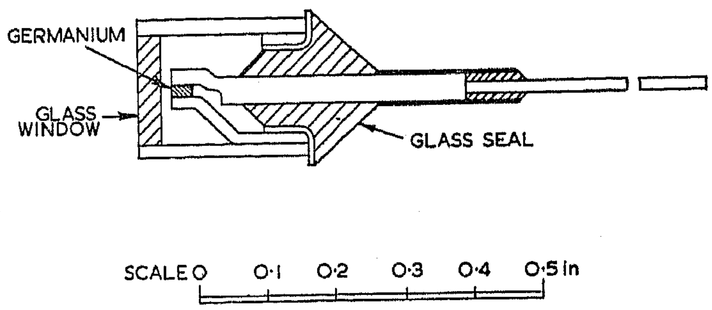


Fig. 2.—Mechanical construction of P50A photocell.

of well-tried characteristics. The n side of the junction is connected to the central lead wire, whilst the p side is connected by a brass strap of high thermal conductivity to the outer case, which consists of a tube with a glass window sealed in the end. Final sealing in a dry atmosphere is achieved by soldering the outer case to the mount carrying the germanium element.

When operation at maximum ratings is intended, advantage may be taken of the good thermal contact between the germanium and the case by mounting the latter in a block of metal as a heat sink. An improvement can also be made by clamping the lead wire in a similar manner.

(3) THE DIFFUSED-IMPURITY JUNCTION

(3.1) Manufacture

The diffused-impurity type of junction was adopted for these photocells to meet the requirements of electrical quality and

large-scale production. This method of junction making[11] has the advantage of producing uniform plane junctions of any desired dimensions, and at the same time it enables practically the whole of a single crystal ingot of germanium to be used for junction making. The ingot, suitably doped to be, say, n type, is cut into slices, and one face of each slice is coated with a suitable acceptor impurity—in the case of these photocells, it is gold. This impurity is diffused into the slice by heat treatment until the desired depth is reached; connection is made to the two sides of the junction by electroplating and/or soldering. The completed slice may be cut into pieces of any desired shape or size. The cutting process leads to strains in the germanium near the surface which reduce the hole lifetime locally and cause excessive reverse currents; this damaged material must be removed by etching.

(3.2) Noise Level

The dark currents of photo-electric cells made by this technique are not quite as low as in the best grown junction types which have been described,[5] but they are very much lower than in the point-contact type. In particular, the noise component of the dark current is much lower than in either type. The noise produced by some typical production photocells when biased in the reverse direction was measured by feeding the noise through an amplifier to a wave analyser, which had an effective bandwidth of 4c/s. The noise could thus be measured as a function of frequency over the a.f. range. The noise power in a typical sample increases with decreasing frequency (Fig. 3), and has the approximate form

$$dN = Af^{-1\cdot 5}\, df$$

The noise currents of many of the cells were of the order of two or three micromicroamperes in a 1c/s band at 1 000c/s—

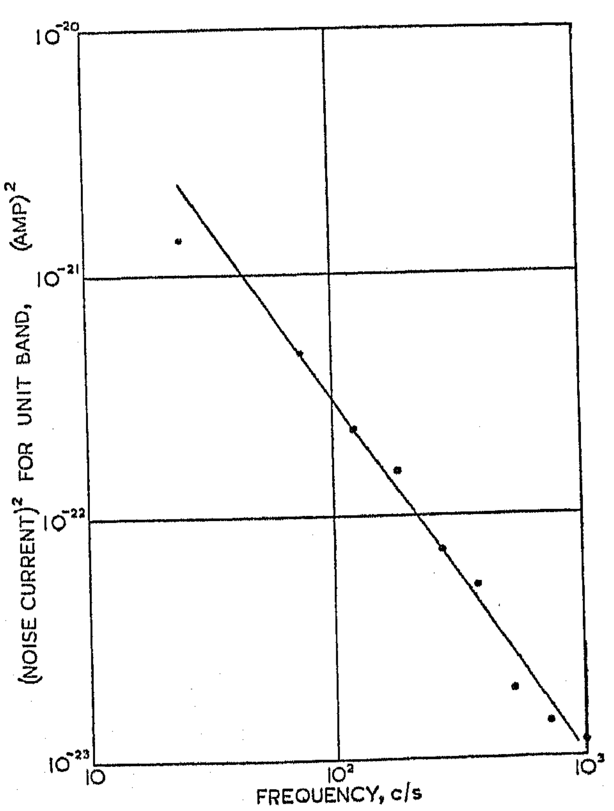


Fig. 3.—Graph of noise current versus frequency for typical P50A photocell.

an order of magnitude lower than those described by Shive. A number of cells were found in which the noise level at 1 000c/s, and at lower frequencies, was lower than this and also lower than the residual noise in the apparatus, and hence it could not be measured. In these cases the noise in the photocell must be almost entirely shot noise. A few cells were found with higher noise levels, and the reason for this variation is not known.

This reduced noise level should enable a significant improvement in overall performance to be obtained in those applications where noise is the limiting factor, which is usually the case in the detection of low modulated light levels. As a result of the low noise level and high sensitivity, the performance of these photocells approaches that of photomultiplier tubes, without the attendant complications of varying sensitivity and elaborate power supplies, so that the apparatus may readily be made portable.

(3.3) Optical Properties

The quantum efficiency of these cells approaches unity in the best specimens, and this results in a high sensitivity. Unfortunately, germanium has a high refractive index, and approximately half the incident light is lost by reflection. The cell responds to light of all wavelengths up to approximately 2 microns, as shown in Fig. 4. Included in this Figure is an

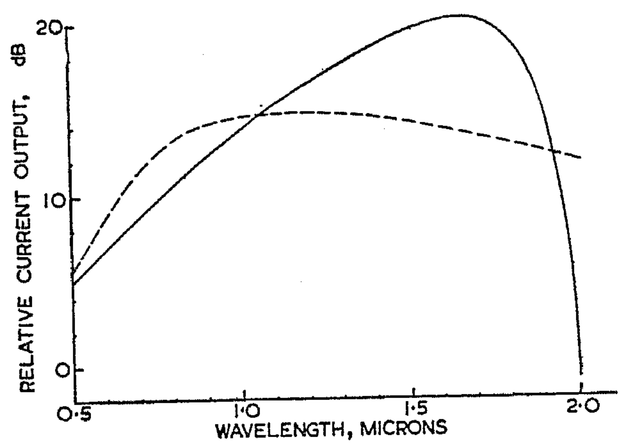


Fig. 4.—Graph of the spectral response of P50A photocell to incident light of uniform intensity plotted against wavelength.
– – – – – Energy content versus wavelength for tungsten-filament lamp at 2 700°K.

output curve for a tungsten-filament lamp, and it will be seen that the photocell responds to much more of the output of the lamp than a photocell which has a long-wave cut-off at, say, 1 micron. The sensitivity of the cell to a given amount of light from a tungsten-filament lamp is thus much higher.

The response of the cell cuts off quite sharply at 2·0 microns. The form of this is in accordance with theoretical predictions and previous experience, but the actual wavelength at which the cut-off occurs is not so. Absorption and photo-conductivity in germanium both cease effectively at 1·8 microns. The photosensitivity of a germanium photocell would be expected to behave in a similar manner, and in fact does so for those samples previously reported.[6] The spectral response of the present photocells was measured by passing light from a tungsten-filament lamp through a quartz-prism monochromator and placing the photocell in front of the exit slit. The variations in output of the lamp and of dispersion of the spectrometer with wavelength were allowed for by making a "dummy run," with a thermopile as the detector. The anomalous long-wave cut-off can be readily demonstrated by carrying out this measurement with a slice of n-type germanium covering the exit slit of the monochromator. This absorbs all light up to a wavelength of 1·8 microns and then becomes transparent. The output of the photocell now rises from zero at 1·8 microns to a peak at 1·9 microns, and then falls to zero again at 2·0 microns. When another make of germanium photocell is used in the experiment, there is no output at any wavelength.

This anomaly is very surprising, since the long-wave cut-off is a function of the energy-band structure and most of the

sensitivity of these photocells resides in the n region, in which the donor is either arsenic or antimony; these produce donor levels so close to the conduction band that they have little effect on the long-wave cut-off, which is therefore at 1·8 micron (e.g. as in the germanium filter experiment described above). However, the acceptor element used in these cells is gold, and this produces an acceptor level 0·13 eV above the valence band.[12] This leaves an energy gap to the conduction band of approximately 0·60 eV, which corresponds to a cut-off of 2·0 microns. But if this explanation is adopted, it must be presumed that the normal photo-conductive mechanism in the n region is somehow suppressed, since otherwise we would expect a discontinuity at 1·8 microns. Measurements have therefore been made both of the distribution of sensitivity across the junction and of the spectral response of the n and p regions separately, using a very small spot of light. The former shows that most of the sensitivity is in the n region and the latter shows no significant difference between n and p regions (Fig. 5). Slices of n-type germanium have been subjected to all the processes to which a complete photocell is subjected, and they showed no apparent anomaly in the long-wave absorption edge.

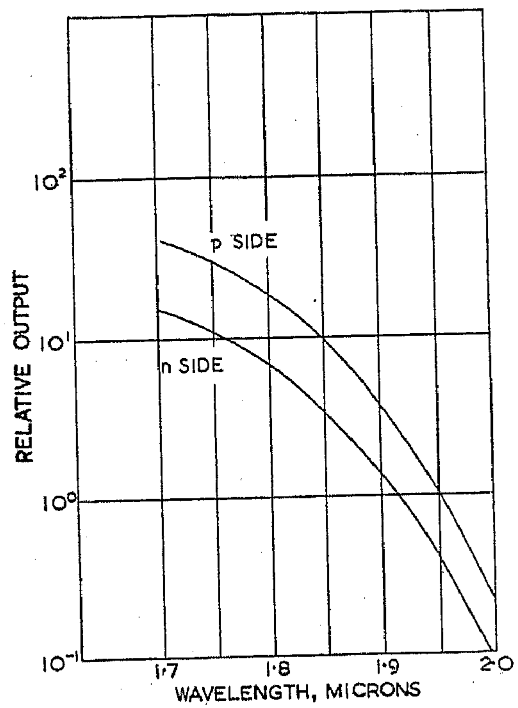


Fig. 5.—Relative output of a photocell when a very small spot of light is directed on to a point in the n region or the p region. This does not show the output of the whole of the n or p regions.

The number of photons in a given amount of radiation is inversely proportional to wavelength, and this would be expected to cause a fall in sensitivity on the short-wave side of 6 dB per octave. The actual fall is more than this, and the reason is not clear, since the common explanation that it is due to increasing absorption in the germanium cannot apply to the geometry used here.

(3.4) Frequency Response

The frequency response of these photocells, like that of junction transistors, is determined by a number of complex factors. The simplest of these is the barrier capacitance, which has been measured for a number of cells as a function of reverse bias (Fig. 6). The capacitance is relatively independent of bias:

$$C \propto V^{-1/3}$$

This is in accordance with the predictions of the theory for this type of p–n junction.[9] The mounting of the photocell itself contributes an additional fixed capacitance of approximately 1·2 $\mu\mu$F.

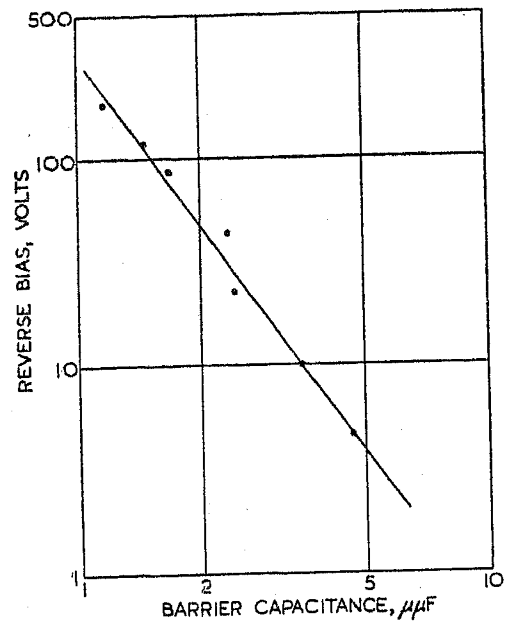


Fig. 6.—Graph of barrier capacitance versus reverse bias for typical P50A photocell at 1 500 c/s.

The extent to which the effective photocell capacitance limits the response depends on the load resistance. With typical values of 5 $\mu\mu$F and 100 kilohms, the 3 dB point is in the region of 350 kc/s. However, other factors may produce an effect at lower frequencies, the most important of which is that, when the cell is illuminated, the minority carriers are generated instantaneously, but they are at various distances from the junction and reach it by diffusion. This introduces a variable time delay, which is due in part to the diffusion process and also on how far away the original carriers were. This cannot be much more than one diffusion length in the cells described, and so the maximum delay is of the order of one lifetime. The lifetime in the n layer is longer than that in the p layer, and has been deduced to be about 15 microsec from measurements of the sensitivity profile. On this basis the response should begin to fall off somewhere above 100 kc/s.

Measurements using mechanical methods of modulating a light beam have shown no fall in output up to 20 kc/s. Above this frequency, some measurements have been made using a flying-spot scanner tube as a modulated light source, and these showed a slight fall in output in the region[13] of 200–300 kc/s.

(4) APPLICATIONS

The high sensitivity and large output characteristics of these devices enable many applications of photocells to be carried out using simpler or smaller and more portable equipment. They are particularly suitable for many industrial "on-off" applications. On the other hand, the combined low noise and good frequency response characteristics make these cells suitable for detecting light beams modulated with a.f. signals, e.g. in sound-film projectors and facsimile transmitters, where the extra sensitivity should allow a stage of amplification to be dispensed with.

An interesting application is the use of infra-red light beams for communication in the field and at sea. In this case the low noise level makes a considerable increase in range possible. At present this development is handicapped by lack of a suitable high-level infra-red modulator. Were one available, a range of some tens of miles would be possible using apparatus of quite small dimensions. Lehovec has suggested that the phenomenon of light absorption by injected carriers in a semi-conductor should be used as the basis of a light modulator,[14] and this has been

demonstrated in practice.[15] If germanium is used as the modulator, light of wavelength longer than 1·8 microns must be used, otherwise the absorption desired will be masked by the main germanium absorption band. Thus normal germanium cells cannot be used as the detector. Lehovec suggests the use of a germanium photocell with a silicon modulator, which would be more difficult to make, while Gibson[15] used a lead-sulphide detector with a germanium modulator. But if a wavelength in the region of 1·9 microns and a germanium modulator were used, it would be within the range of sensitivity of the present photocell and such a system should have a considerably improved overall performance. Unfortunately, this type of modulator cannot easily be used at high light levels.

In some applications of this type wide-band modulation is not required, as for telegraphy and for recording the interruption of a beam of light by the movement of persons or things, e.g. the movement of aircraft on a large airfield. In these cases a simple mechanical modulator can be used, such as a vibrating shutter or mirror, and long ranges can be obtained with very small portable equipment having only a small power consumption.

Fig. 7 shows the output characteristics of the P50A photocell.

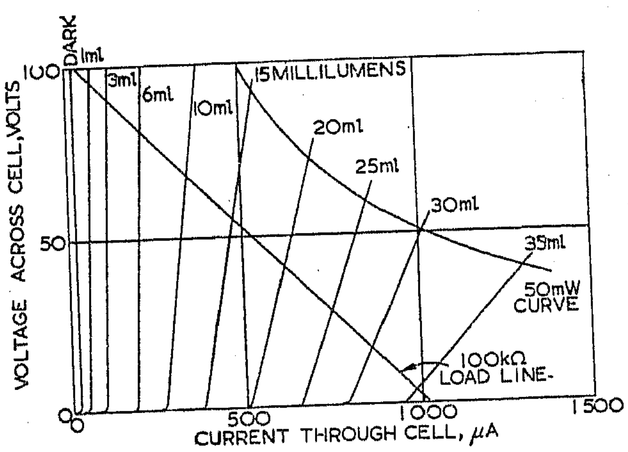


Fig. 7.—Graph of current versus applied voltage for various light levels at a temperature of 20°C for a typical P50A photocell.

The similarity to the collector characteristic of a junction transistor should be noted. A great many of the applications of this photocell are of the large-signal "on-off" type, and in these it is often necessary to obtain the maximum output power to operate a relay or to trigger a cold-cathode discharge tube. The optimum load is of the order of 100 000 ohms, and the power available is about 100 mW if the dissipation of the cell is kept below the maximum of 50 mW. This load resistance is much higher than can readily be obtained in relay coils, and advantage may be taken of the high forward conductance of the photocell to reduce the optimum load resistance to a conveniently low value. At the same time, the available power can be increased to approximately 1 watt. The circuit (Fig. 8) uses a single-coil magnetic amplifier of a type due to Ramey.[16,17] The core of this coil is of Permalloy F or similar material having a rectangular hysteresis loop, and therefore has a sharp transition from the unsaturated to the saturated state. The circuit is so designed that the core does not quite reach saturation when the full alternating voltage is applied; under these conditions a magnetizing current of approximately 1 mA flows. When the photocell is fully illuminated, it will pass this current in either direction with negligible voltage drop. When the illumination is removed, the photocell becomes a rectifier and will only pass current in one direction. The magnetic core becomes saturated and the inductive reactance of the coil is removed from the circuit, leaving only its direct resistance, which can be made very much smaller. The remainder of the circuit now operates as a half-wave rectifier and load resistance, and with a load of 1 000 ohms nearly 1 watt can be obtained. If the load is inductive it is necessary to shunt it with a rectifier in order to short-circuit the

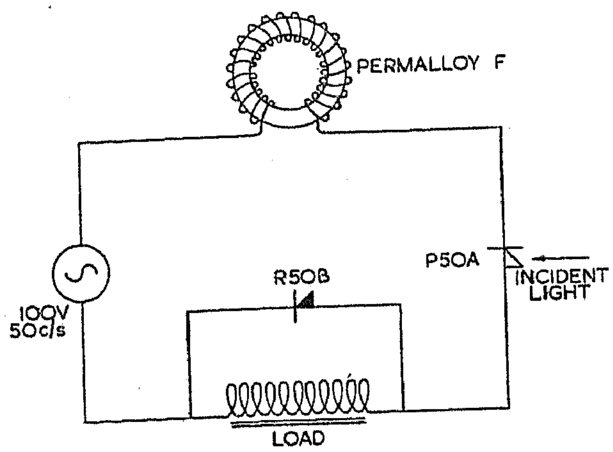


Fig. 8.—Simple magnetic-amplifier circuit for use with junction photocell.

inductance during the half-cycle when the photocell is not conducting. It should be noted that this circuit is linear in the sense that intermediate values of mean output current can be obtained by intermediate degrees of illumination.

An important field of application for this type of photocell is in the reading of holes in punched-card systems and teleprinter tape. In these applications minimum dimensions are essential, and the P40A cell has therefore been developed, with an overall diameter of only 0·087 in. The size of the germanium element in this is only 0·035 in × 0·035 in × 0·020 in, and the other dimensions are correspondingly smaller. This has resulted in a much smaller cooling capacity (m value), but nevertheless this photocell shares most of the advantages of the larger P50A type. It has complete hermetic sealing and a large power output; for example, a 30-volt signal can be obtained to work a cold-cathode tube.

(5) CONCLUSIONS

The diffused-impurity junction photocell has the desirable properties of small size, large output, good sensitivity and stable characteristics. The noise level in many samples has been found to be remarkably low. It has been shown that temperature rise rather than absolute temperature is the important criterion in determining the maximum ratings under given conditions.

The spectral response of these cells differs from that of previously described types in having a somewhat extended long-wavelength cut-off. No satisfactory theoretical explanation of this phenomenon has been given. A suggested application of this effect is described, together with a number of other applications, including the use of a simple magnetic amplifier which makes use of the high forward conductance of the photocell.

(6) REFERENCES

(1) Smith, W.: "Effect of Light on Selenium During the Passage of an Electric Current," *Nature*, 1873, **7**, p. 303.
(2) Smith, R. A.: "Infra-red Photo-conductors," *Advances in Physics*, 1953, **2**, p. 321.
(3) Veszi, G. A.: "The Selenium Barrier Layer Photocell," *Electronic Engineering*, 1941, **14**, p. 436.
(4) Chapin, D. M., Fuller, C. S., and Pearson, G. L.: "A New Silicon *p–n* Junction Photocell for Converting Solar Radiation into Electrical Power," *Journal of Applied Physics*, 1954, **25**, p. 676.

(5) SHIVE, J. N.: "Properties of the M.1740 *p–n* Junction Photocell," *Proceedings of the Institute of Radio Engineers*, 1952, **40**, p. 1410.

(6) SHIVE, J. N.: "The Properties of Germanium Photo-Transistors," *Journal of the Optical Society of America*, 1953, **43**, p. 239.

(7) PIETENPOL, W. J.: "*p–n* Junction Rectifier and Photocell," *Physical Review*, 1951, **82**, p. 120.

(8) BILLIG, E.: "Application of the Image-Force Model to the Theory of Contact Rectification and of Rectifier Break-down," *Proceedings of the Royal Society*, 1951, **207**A, p. 156.

(9) SHOCKLEY, W.: "The Theory of *p–n* Junctions in Semi-conductors and *p–n* Junction Transistors," *Bell System Technical Journal*, 1949, **28**, p. 435.

(10) HALL, R. N.: "Power Rectifiers and Transistors," *Proceedings of the Institute of Radio Engineers*, 1952, **40**, p. 1512.

(11) DUNLAP, W. C.: "Diffusion of Impurities in Germanium," *Physical Review*, 1954, **94**, p. 1531.

(12) DUNLAP, W. C.: "Electrical Properties of Gold Germanium Alloys," *ibid.*, 1953, **91**, p. 1282.

(13) TOMBS, D. M., and TAYLOR, J. E.: Private Communication.

(14) LEHOVEC, K.: "New Photoelectric Devices Utilizing Carrier Injection," *Proceedings of the Institute of Radio Engineers*, 1952, **40**, p. 1407.

(15) GIBSON, A. F.: "Germanium Modulator for Infra-red Communication," *Electronics*, 1954, **27**, No. 10, p. 155.

(16) RAMEY, R. A.: "On the Mechanics of Magnetic Amplifier Operation," *Transactions of the American I.E.E.*, 1951, **70**, Part II, p. 1214.

(17) SPENCER, R. H.: "Transistor Controlled Magnetic Amplifier," *Electronics*, 1953, **26**, No. 8, p. 136.

(18) SHOCKLEY, W.: "Electrons and Holes in Semi-conductors" (D. Van Nostrand Co., 1950), p. 112.

[The discussion on the above paper will be found on page 786.]

621.375.4 : 621.314.7

Paper No. 1861 R
Apr. 1955

TRANSISTOR POWER AMPLIFIERS

By R. A. HILBOURNE, B.Sc., and D. D. JONES, M.Sc.

(*The paper was first received* 4*th March, and in revised form* 18*th March*, 1955. *It was published in April*, 1955, *and was read before a* JOINT MEETING *of the* RADIO *and* MEASUREMENTS SECTIONS 11*th May*, 1955.)

SUMMARY

The factors governing the power-handling capabilities of *p–n–p* junction transistors in sinusoidal amplifiers are discussed. It is shown that, although the maximum allowed collector dissipation, $P_{c\,max}$, is an important transistor design criterion, it does not necessarily indicate the limit to the maximum output power possible; limitations set by other factors, notably the variation of current gain with emitter current, become particularly significant in Class B amplifiers because of their very high efficiencies.

The relative importance of Class A and Class B amplifiers is considered, and the three possible Class B push-pull arrangements are analysed in detail. It is shown that load-line techniques, such as are used in designing thermionic-valve power amplifiers, are not really suitable for transistors, and a different approach is made.

It is concluded that the common-collector Class B push-pull amplifier has a number of attractive features which favour its use. The design of a driver stage for this amplifier is described.

Complementary arrangements of *n–p–n* and *p–n–p* transistors are examined, but it is found that their advantages are not very great.

Finally, some non-sinusoidal power-amplifier applications are mentioned, and it is seen that transistors can be very important in this context.

LIST OF PRINCIPAL SYMBOLS

A = Current (or voltage) gain at the frequency of the fundamental component of the output signal.

α = Emitter-to-collector current gain, di_c/di_e, under specified conditions of current and voltage bias.

α_0 = Approximate value of α when the emitter-current bias is zero.

α_{cb} = Base-to-collector current gain, di_c/di_b, and is related to α by $\dfrac{\alpha}{1-\alpha}$.

$\alpha_{cb0} = \dfrac{\alpha_0}{1-\alpha_0}$.

B = Fractional third-harmonic content.

β = Factor describing the variation of current gain, α, with emitter current.

G = An index of gain given by the ratio P_{load}/P_{gen}.

i_e = Current flowing in the emitter terminal of the transistor.

i_c = Current flowing in the collector terminal of the transistor.

i_b = Current flowing in the base terminal of the transistor.

i_{es}, i_{cs}, i_{bs} = Alternating signal currents flowing in the emitter, collector and base terminals.

$I_{c\,max}$ = Maximum allowed value of i_c, determined by distortion considerations.

M = Figure of merit for a power transistor = $I_{c\,max}\,V_{c\,max}$.

$P_{c\,max}$ = Maximum allowed value of mean collector dissipation.

P_{gen} = Total power developed in the output circuit of the driver stage.

P_{load} = Power into the load.

P_{gain} = Power gain of the amplifier.

P_{lim} = Maximum value of output power, the limit being determined by M.

R_L = Load resistance.

r_e = Emitter resistance as defined in the small-signal T-equivalent network.

r_b = Base resistance defined in the same way as r_e.

r_c = Collector resistance defined in the same way as r_e.

r_s = Effective source resistance.

V_{bias} = Forward bias voltage applied to overcome cross-over distortion.

V_c = Collector voltage.

$V_{c\,max}$ = Maximum allowed value of V_c.

v_L = Alternating voltage appearing across the amplifier load.

v_s = Effective input signal from the generator.

(1) INTRODUCTION

Many applications where transistors can be used to advantage require that large output powers be handled. For such requirements the frequencies involved are generally in the audio range (up to 10kc/s). When handling large output powers transistors have certain advantages over other amplifying devices; thus they can operate from low-voltage power supplies (e.g. 12 or 24 volts d.c.) and are extremely efficient. This high overall efficiency is due to the fact that the collector characteristics bottom at practically zero voltage (less than 0·1 volt) and cut off at substantially zero current (usually less than 0·1 mA). The low-voltage operation is not always an advantage because it necessitates a large output current which may lead to undesirable distortion. A distinct limitation, particularly in the case of germanium transistors, is that the transistor characteristics deteriorate when either the ambient temperature and/or collector dissipation cause the collector-junction temperature to exceed a certain value (this limit is usually less than 100° C).

The object of the paper is to examine the performance of low-power *p–n–p* junction transistors as power amplifiers. Although the output powers considered are rather small (e.g. 0·5 watt) the principles established are quite general and can be readily applied to higher powers. The variations and limitations of transistor parameters are considered in relation to power amplification, and the results are applied to actual circuit design; in particular, the advantages of the various transistor circuit arrangements are discussed with special emphasis placed on distortion.

Although a large number of papers have appeared on various aspects of transistor circuits, only a few have dealt with power amplifiers. Unfortunately these give little detailed consideration to problems of distortion. Shea[4] considers in detail the distortion in a common-base Class A amplifier due to variation of emitter resistance with emitter-current bias. This distortion can be more easily overcome in practice than the distortion due to variation in the current-gain factor. The latter variation is of great importance in the design of transistors capable of handling large output powers.

From a consideration of the possible circuits it is found that, for a given output power, high power-gain is, in general, accompanied by a high level of distortion. The power gains with the

The paper is a communication from the Staff of the Research Laboratories of The General Electric Company, Limited, Wembley, England.

different circuit arrangements decrease in the order: common emitter, common collector, common base. As a good practical compromise between gain and distortion it is concluded that the common-collector arrangement is preferred; the design of a suitable driver stage for this arrangement is discussed. The primary cause of distortion is the variation of current gain with emitter current.

Sziklai[3] has discussed the possibilities of using parallel arrangements of *n–p–n* and *p–n–p* transistors in power amplifiers. This complementary symmetry approach, although very interesting and in some cases advantageous, is discussed only briefly because of the general lack of *p–n–p* and *n–p–n* transistors exhibiting the same variation of current gain with emitter current.

Frequency limitations of power amplifiers are not discussed in the present paper. In practice these effects are not serious in the frequency range concerned.

(2) TYPES OF POWER AMPLIFIERS

For power amplification of audio-frequency sinusoidal waveforms two basic arrangements, known as Class A and Class B circuits, are generally used.[5]

In the Class A amplifier the transistor is biased so that neither the collector current nor voltage is cut off during any part of the input waveform. Under these conditions the mean collector bias is unaffected by the signal.

The maximum possible efficiency, η, of a Class A amplifier for a sinusoidal input is 50%, where η is defined by

$$\eta = \frac{\text{Alternating output power into the amplifier load}}{\text{Total alternating and direct power applied to the amplifier}}$$

In practice, maximum efficiency is not obtained because the current-gain variations limit the maximum collector current amplitude. However, efficiencies of about 48% are possible using junction transistors.

In the Class B amplifier the transistor is biased so that amplification occurs only over one half-cycle of the input waveform. That is, in the quiescent state (in the absence of input signal) the transistor is cut off and the power dissipated in it is approximately zero. This cut-off condition can take two forms. In the first the transistor is biased to heavy collector current at very low collector voltage (point A in Fig. 1) in the quiescent condition, the transistor then (in the common-base arrangement) amplifying the negative half-cycle of the input waveform. In the second the transistor is biased to low collector current, high collector voltage (point B in Fig. 1) in the quiescent condition, the transistor then amplifying the positive half-cycle of the input waveform. The first method is in practice extremely inefficient but has certain advantages when using point-contact transistors.[6] For junction transistors the second method is almost invariably used.

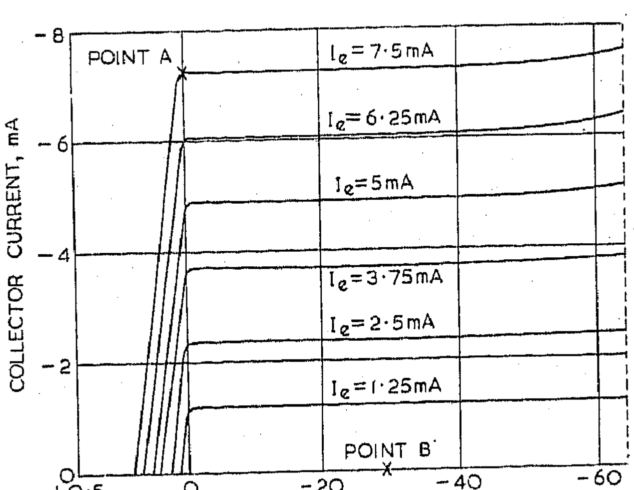


Fig. 1.—Typical static collector characteristics of an EW52 junction transistor in the common-base connection.

Because a transistor biased for Class B operation amplifies only one-half of the input signal, it is necessary to use two transistors in push-pull as described later, in order to obtain a sinusoidal output waveform.

The maximum theoretical efficiency of a Class B push-pull amplifier[5] is 78·5% when amplifying a sinusoidal waveform. Using junction transistors efficiencies of 70% and greater have been obtained.

In practice the Class B amplifier, in the strict sense, is seldom used. The emitter is generally biased slightly forward in the quiescent condition in order to overcome an effect known as cross-over distortion. This reduces the overall efficiency, but values of the order of 65–70% are still obtained.

(3) LIMITATIONS OF POWER AMPLIFIERS

The ultimate limit to the power a transistor amplifier can handle is set by the maximum temperature at which the collector junction can be operated. For most transistors this limit is usually below 100°C. The amount of power that can be dissipated in the collector before this limit is reached depends on the ambient temperature and on the cooling system used, but can be determined from two simple experiments.

At a given collector voltage the collector current that flows when the emitter is open-circuit (generally known as I_{co}) increases exponentially with temperature. Typically, it approximately doubles for every 10°C rise in junction temperature. The collector may therefore be used as its own thermometer and can be calibrated by heating in an oven and measuring I_{co} at, say, −6 volts at various known temperatures.

In the next experiment the transistor, fitted with the particular cooling arrangement to be used, is operated at various values of collector dissipation (established by varying the emitter current). The dissipation is maintained for a time long in comparison with the thermal time-constant of the transistor with its cooling arrangement (this time-constant may be of the order of minutes). The dissipation is then suddenly removed and I_{co} when V_c is −6 volts is immediately measured. The collector-junction temperature may then be read off the calibration chart obtained from the first experiment. For example, with a low-power junction transistor the junction temperature increased by 1°C every 5mW rise in dissipation, whereas for a medium-power transistor the figure was 1°C per 40mW.

For a given ambient temperature the maximum collector power dissipation $P_{c\,max}$ can thus be obtained. So long as the time-constant of the signal being amplified is much shorter than the thermal time-constant of the transistor (and this generally holds for most audio-amplifier applications) the value of dissipation to be taken into account is the mean rather than the peak value.

In practice, factors other than $P_{c\,max}$ may determine the maximum output power that can be obtained. However, even then it is desirable to use as much transistor cooling as possible in order to keep the collector-junction temperature to a minimum, thereby keeping the effects of high temperature on transistor characteristics as small as possible.

One factor that limits transistor operation is the maximum voltage $V_{c\,max}$ that may be applied to the collector. It is important to realize that transistor amplifiers in all possible configurations (common base, common emitter or common collector) are cut off by the common base I_{co} characteristic, i.e. when the emitter current is zero. The peak voltage that can be applied to the collector is a voltage slightly below that at which the I_{co} characteristic starts "softening" rapidly. Families of characteristics for a common-emitter arrangement, for example,

corresponding to fixed values of base current are of little value unless they are plotted for both positive and negative base current (Fig. 2). In Class B amplifiers the collector voltage will generally rise to twice the battery supply voltage so that the latter is limited to $0 \cdot 5 V_{c\,max}$.

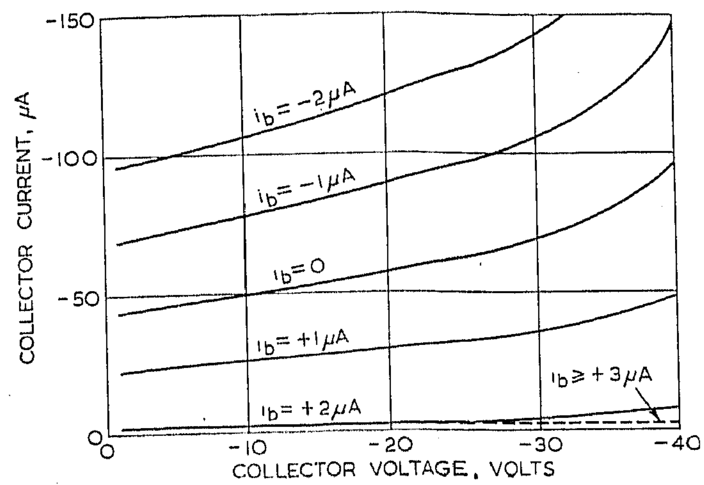


Fig. 2.—Static collector characteristics of an EW52 junction transistor for small values of positive and negative base current (common emitter).

Another serious limitation is often set by $I_{c\,max}$, the maximum collector current that can be used. This limit is usually set by the amount of distortion that can be tolerated. Fig. 3 shows the variation of α, the current gain (common-base operation), with emitter current for a junction transistor.

Thus it is seen that α increases at first with increasing emitter current and then drops approximately linearly. An approximation to the variation shown in Fig. 3 may be written

$$\alpha = \alpha_0 - \beta i_e \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

This relationship is also shown in Fig. 3.

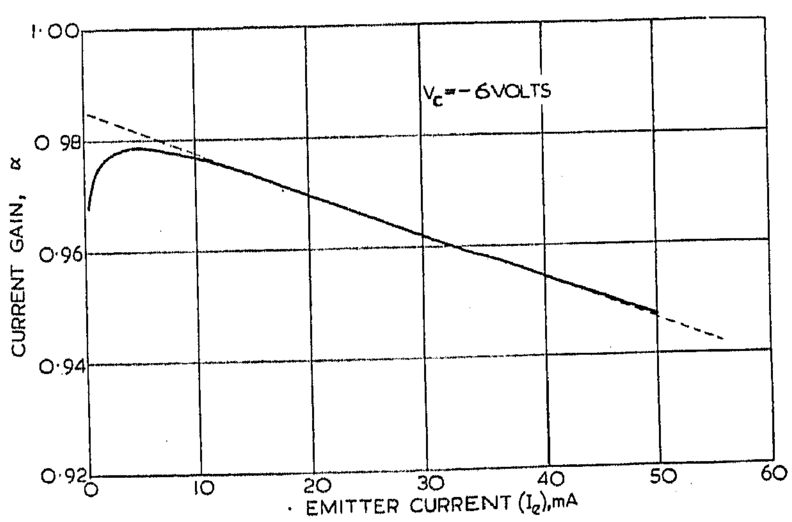


Fig. 3.—Variation of current gain α with emitter current for a typical EW52 junction transistor.

—— Experimental.
---- Theoretical (based on $\alpha = \alpha_0 - \beta I_e$, where $\alpha_0 = 0 \cdot 985$ and $\beta = 0 \cdot 00077$/mA).

In the common-emitter arrangement the current gain factor becomes $\alpha/(1 - \alpha)$ and the variation of this with emitter current, which is much more serious than that of α, is shown in Fig. 4.

The variation of α with emitter current has been discussed by Webster.[7] Briefly, the initial rise in the characteristic at low values of emitter current depends on surface recombination effects. In a *p–n–p* transistor the emitter current consists mainly of positive holes flowing from the emitter to the base region and partly of electrons flowing from the base to the emitter. A number of the holes migrate to the germanium surface in the vicinity of the emitter-base junction and are lost there by

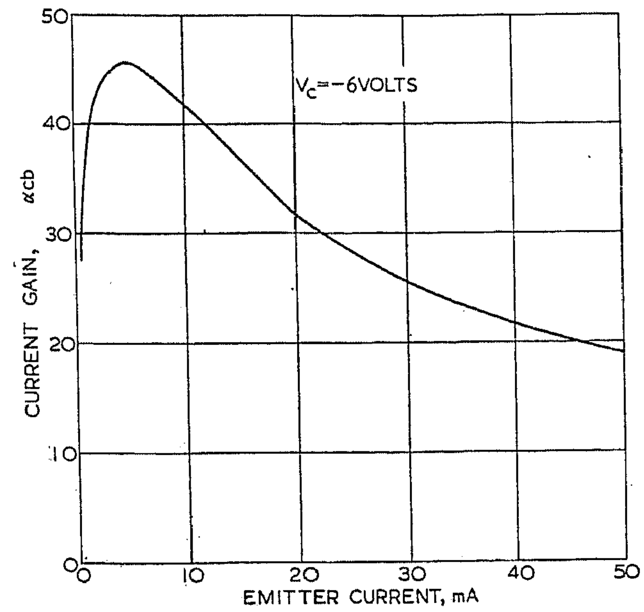


Fig. 4.—Variation of current gain α_{cb} [$=\alpha/(1-\alpha)$] with emitter current for a typical EW52 junction transistor.

recombination with electrons. This has the effect of reducing the current gain, α, which is unity when the emitter current is entirely composed of holes and all these reach the collector. As the emitter current increases, surface recombination has a lower percentage effect on current gain, which thus rises. Increasing surface recombination tends to flatten out the initial peak of the α_{cb}/I_e curve (Fig. 4).

As the emitter current becomes very large the high density of holes in the *n*-type base region gives rise to a large diffusion gradient between emitter and collector. This results in the density of holes in the *n* region being very high near the emitter. In attempting to maintain the base region in a neutral state a corresponding electron density is set up. This electron density is considerably higher than that existing in the *n* region at low emitter currents, and can be provided only by drawing current from the base electrode. This causes a reduction in current gain because some of the emitter current now consists of electron current from the base region. The magnitude of this reduction depends on the actual current density at the emitter-base junction and can be reduced by increasing the emitter area.

This variation of current gain with emitter current gives rise to distortion, and if a maximum limit is set to this, the emitter current must be maintained below a certain maximum value. Since the collector current does not differ very appreciably from emitter current, this also limits the collector current to a maximum value $I_{c\,max}$ for a given amplifier arrangement.

A useful guide to the performance of a transistor in a power amplifier is the figure of merit M described by the relationship

$$M = V_{c\,max} \times I_{c\,max}{}^{(1)}$$

This is sometimes useful in assessing both the power capabilities of the transistor and also the minimum amount of cooling necessary to take full advantage of the transistor.

Another limitation is set by the variation of r_e, the emitter resistance of the small-signal T-equivalent network, with emitter current. This is shown in Fig. 5. At a temperature of 20°C the variation is given approximately by the expression

$$r_e = \frac{25}{i_e}$$

where i_e is the emitter current in milliamperes and r_e is in ohms. At the beginning of the half-cycle being amplified in a Class B amplifier r_e is very high but decreases rapidly with signal amplitude. This variation gives rise to cross-over distortion (Fig. 7).

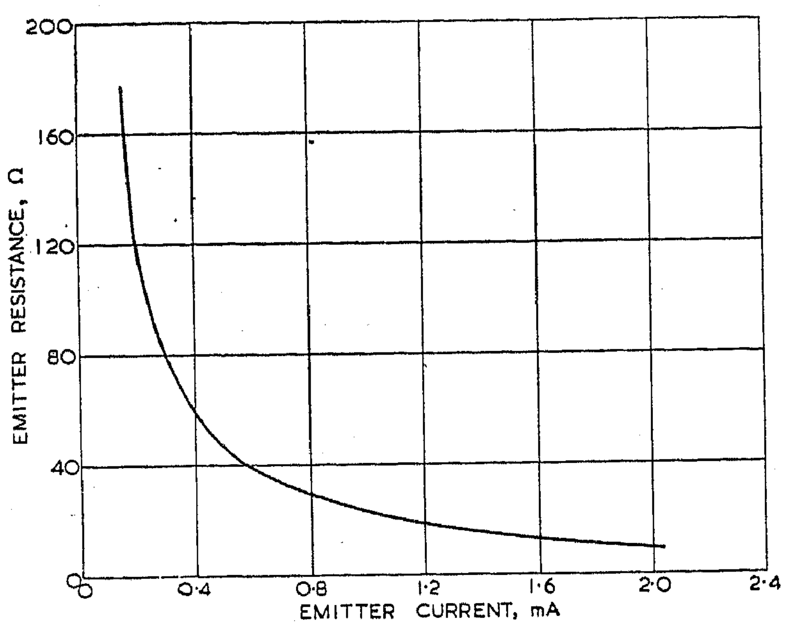


Fig. 5.—Variation of emitter resistance r_e with emitter current for a typical EW52 junction transistor.

Once its significance is appreciated it is relatively simple to design a circuit to correct for it.

(4) RELATIVE IMPORTANCE OF CLASS A AND CLASS B POWER AMPLIFIERS WHEN USING TRANSISTORS

The chief advantages of the Class A amplifier are: first, that only a single transistor is used, and secondly, that no cross-over distortion occurs. It can, however, suffer from considerable second-harmonic distortion.[4] Maximum power is dissipated in the transistor at zero-input-signal conditions, and thus the maximum possible output power is $\frac{1}{2}P_{c\,max}$.

At zero input signal the dissipation of the Class B amplifier is almost zero. Assuming an overall efficiency of 67% the maximum possible output is $\frac{67}{100-67} \times P_{c\,max}$, i.e. twice $P_{c\,max}$. This can, of course, be obtained only if the time-constant of the signal is considerably less than the thermal time-constant of the transistor assembly. The push-pull operation reduces even-harmonic distortion to a negligible amount.

For transistors having the same basic electrical design, M will be approximately the same for both Class A and Class B operations in a given configuration (e.g. common base). Thus the maximum power output P_{lim} per transistor at a given distortion level will be of the same order for both Class A and B arrangements. For the Class A arrangement $P_{c\,max}$, the maximum power that can be dissipated in the transistor, must be equal to twice P_{lim} to use the transistor to its electrical performance limit. However, for the one transistor in a Class B arrangement the value of $P_{c\,max}$ is one-half the value of P_{lim}. Thus the degree of cooling required in the Class B case is considerably lower than for the Class A. Thus, bearing in mind the importance of adequate cooling of transistors, the Class B arrangement is in general more important than the Class A.

(5) METHODS OF DESIGN

In the design of small-signal transistor amplifiers it is usual to employ an a.c. equivalent network the parameters of which can be assumed constant over the applied signal swing. In power amplifiers the parameters vary considerably and the chief interest in the small-signal equivalent network is as a means of expressing these variations.

In the case of thermionic-valve power amplifiers it is usual to calculate performance graphically by superimposing a load line on the anode characteristics. These characteristics are independent of the impedance of the source feeding the grid. In the transistor, however, similar load-line techniques, although they have been suggested,[2] are of little value because the collector characteristics are highly dependent on the emitter- (or base-) source impedance. This may readily be seen by considering the case of the small-signal transistor amplifier operated from a source of impedance r_s. Characterizing the amplifier by the general equations[4]

$$\left.\begin{aligned} V_1 &= r_{11}i_1 + r_{12}i_2 \\ V_2 &= r_{21}i_1 + r_{22}i_2 \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad (2)$$

we find that the output impedance is given by

$$R_{out} = r_{22} - \frac{r_{12}r_{21}}{r_{11} + r_s} \quad . \quad . \quad . \quad . \quad . \quad (3)$$

Transistor characteristics are generally plotted for the case when r_s is infinite, the slope of the characteristics passing through the bias point concerned then being given by r_{22}. For a thermionic-valve amplifier r_{12} is zero, whereas for transistors it is usually in the range 50–500 ohms.

In order to use load-line techniques with transistors it is necessary to plot the collector characteristics for fixed values of source voltage v_s, and using the precise value of source resistance r_s that is to be used in the actual circuit. Since this means a new set of characteristics every time a source of different resistance is used it is generally advisable to avoid the use of load-line techniques.

The most important use of static-collector characteristics (for constant value of I_e), such as those shown in Fig. 1, is to determine the maximum value of collector voltage that can be used.

The problem is also simplified by the fact that the load resistance is usually very low (e.g. 100 ohms) and that no attempt is made at impedance matching of the amplifier to the load.

The design of the Class B power amplifier can be approached by assuming first that the transistor characteristics are perfect, i.e. that

(a) The cut-off collector current is zero.
(b) The collector characteristics bottom at zero collector voltage.
(c) The parameters of the small-signal equivalent network are constant over the range of large signal used.
(d) The transistor output resistance is infinite compared with the load.

In practice (a), (b) and (d) are sufficiently true that no further first-order correction need be made on their account. Regarding the third assumption, it is found that emitter resistance r_e and current gain α do have significant variations, and these in fact become the principal factors to be considered in the amplifier design. These effects are dealt with in full in later Sections.

Assuming that (a) and (b) are reasonably true, the output load resistance can be obtained from the output power required and the available h.t. supply voltage. The mean a.c. power developed in a load resistance R_L by a complete sine wave of peak voltage V_p is given by[5]

$$P_{mean} = \frac{V_p^2}{2R_L} \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

For a Class B push-pull power amplifier the peak voltage is equal to the h.t. voltage, hence

$$\text{Mean a.c. power} = \frac{V_{ht}^2}{2R_L} \quad . \quad . \quad . \quad . \quad . \quad (5)$$

where R_L is the load resistance presented to each collector in turn. For example, for 500 mW output with a 20-volt h.t. supply a load resistance of 400 ohms is required.

(6) PERFORMANCE OF THE THREE BASIC CLASS B PUSH-PULL AMPLIFIER ARRANGEMENTS

In this Section the three basic arrangements are considered in turn. A method for overcoming cross-over distortion is discussed. The amount of distortion caused by the variation of current gain with emitter current is estimated so that the three arrangements may be compared. Power gains are also calculated.

(6.1) Common-Base Amplifier

The basic circuit of the common-base push-pull amplifier is shown in Fig. 6. The a.c. input signal is fed into the emitter

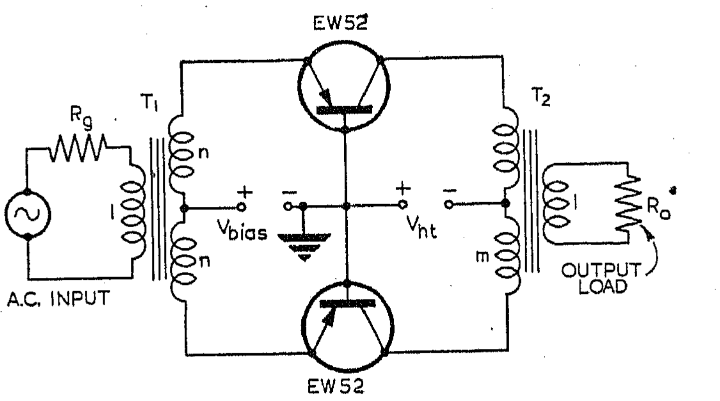


Fig. 6.—Basic circuit of the common-base push-pull amplifier.
$r_S = n^2R_0$
$R_L = m^2R_0$

electrodes of the two transistors via the phase-splitting transformer T_1. The output transformer T_2 is connected between the collector electrodes, with the h.t. supply applied to the centre tap of the primary winding.

Forward emitter bias (V_{bias}) is applied to the centre tap of the phase-splitting transformer T_1. This bias must be fed from a low-resistance source since this is in series with the input signal. This bias generator resistance cannot be decoupled by a capacitor since the diode action of the emitter electrodes rectifies the input signal which would set up an additional direct voltage across the bias resistance.

The small-signal input resistance of a transistor in the common-base arrangement is given[4] by

$$r_{in} = r_e + r_b\left[1 - \frac{\alpha(r_c + r_b)}{r_b + r_c + R_L}\right] \quad . \quad . \quad . \quad (6)$$

or

$$r_{in} \simeq r_e + r_b(1 - \alpha) \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

provided that R_L is small compared with r_c, as is generally the case in power amplifiers.

If, therefore, the emitter is fed from a source resistance r_s (which includes the d.c. bias resistance) and signal voltage v_s, the emitter signal current i_{es} is given by

$$i_{es} = \frac{v_s}{r_{in} + r_s} = \frac{v_s}{r_s + r_e + r_b(1 - \alpha)} \quad . \quad . \quad . \quad (8)$$

Hence the collector signal current i_{cs} is given by

$$i_{cs} = \alpha i_{es} = \frac{\alpha v_s}{r_s + r_e + r_b(1 - \alpha)} \quad . \quad . \quad . \quad (9)$$

but, as shown in Fig. 5, the emitter resistance varies with emitter current and is highest at low values of emitter current causing cross-over distortion. This cross-over distortion can be reduced by increasing r_s and applying forward emitter bias, which reduces the value of r_e in the quiescent state.

These effects are shown in Fig. 7, in which a generator resistance of 100 ohms is used, with a forward bias of 120 mV. This gives a

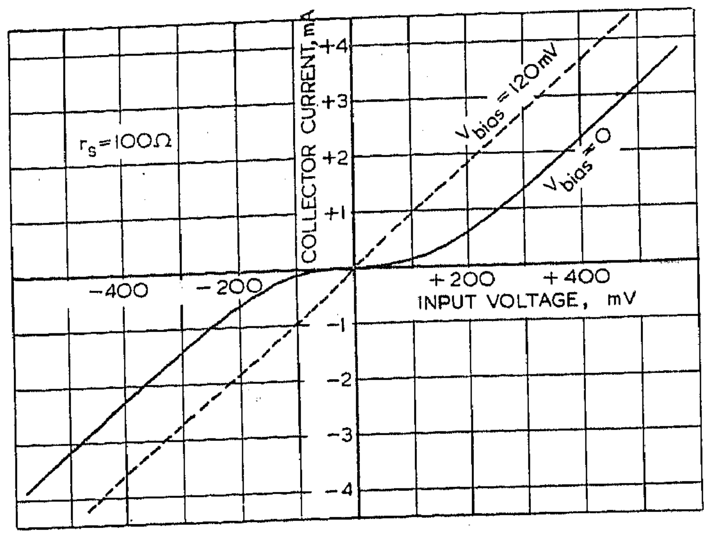


Fig. 7.—Cross-over distortion of a common-base amplifier with a generator resistance of 100 ohms, and the effect of 120 mV forward bias.

standing collector current of 0·33 mA per transistor, and hence a small reduction in the maximum possible efficiency.

With a collector load resistance R_L the voltage gain of the common-base amplifier becomes

$$\frac{v_L}{v_s} = \frac{i_{cs}R_L}{v_s} = \frac{\alpha R_L}{r_s + r_e + r_b(1 - \alpha)} \quad . \quad . \quad . \quad (10)$$

Hence, if r_s is large compared with $r_e + r_b(1 - \alpha)$,

$$\frac{v_L}{v_s} \simeq \frac{\alpha R_L}{r_s} \quad . \quad . \quad . \quad . \quad . \quad . \quad (11)$$

The power into the load resistance is therefore given by

$$P_{load} \simeq \frac{v_L^2}{R_L} \simeq \frac{v_s^2\alpha^2R_L^2}{R_L r_s^2} \simeq \frac{v_s^2\alpha^2R_L}{r_s^2} \quad . \quad . \quad . \quad (12)$$

The maximum available power from the generator is given by

$$P_{source} = v_s^2/4r_s \quad . \quad . \quad . \quad . \quad . \quad . \quad (13)$$

The power gain of the common-base amplifier, therefore, which is defined by the expression

$$\text{Power gain} = \frac{\text{Power delivered to load}}{\text{Maximum power available from the generator}}$$

is given by

$$\text{Power gain} = \frac{v_s^2\alpha^2R_L}{r_s^2}\frac{4r_s^2}{v_s^2} \quad . \quad . \quad . \quad . \quad (14)$$

$$= \frac{4\alpha^2R_L}{r_s} \simeq \frac{4R_L}{r_s} \quad . \quad . \quad . \quad . \quad (15)$$

since α is approximately equal to unity.

As shown in Section 5, R_L is determined by the h.t. supply voltage and the output power required, and has, in general, a low value, particularly at high output power. Thus the common-base amplifier has a low power gain for large output powers, and is more suitable for medium-power amplifiers. For example, if $V_{ht} = 30$ volts, $P_{load} = 200$ mW, and $R_L = 2{\cdot}25$ kilohms, then with $r_s = 100$ ohms, $P_{gain} = 90$.

For this example, using EW52 transistors (see Fig. 5), the variation of current gain α over the range of emitter-current swing is less than 0·5%, and thus the distortion due to this can be neglected. Although the gain of this amplifier is very low at high output levels, its low distortion characteristics are such that it may still be advantageous to use it in applications where low distortion is the sole criterion.

(6.2) Common-Emitter Amplifier

The basic circuit of the common-emitter amplifier is shown in Fig. 8.

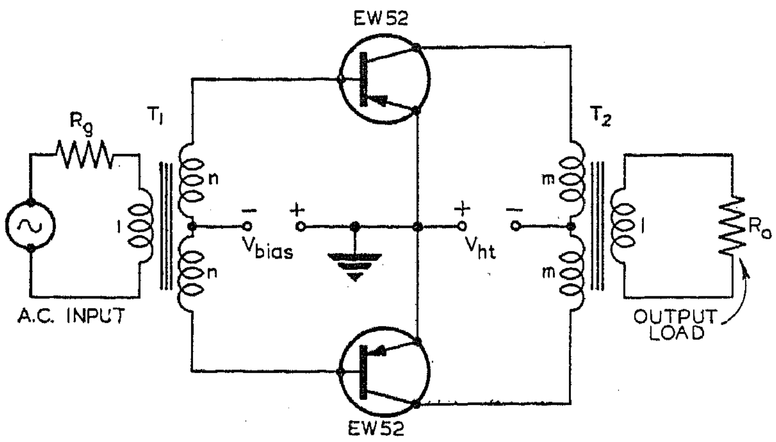


Fig. 8.—Basic circuit of the common-emitter amplifier.

$r_S = n^2 R_g$
$R_L = m^2 R_0$

The input signal is applied between the bases and emitters via a transformer T_1, and the output taken between the collectors and emitters by transformer T_2. As for the common-base connection, forward bias is applied to the input electrode, which in this case is the base and hence the bias is negative.

The input resistance of a transistor in the common-emitter connection at low frequencies is given[4] by

$$r_{in} = r_b + r_e\left(\frac{\dfrac{r_c + R_L}{r_c + r_b}}{\dfrac{R_L + r_e}{r_c + r_b} + (1-\alpha)}\right) \quad . \quad . \quad (16)$$

This may be written approximately as

$$r_{in} = r_b + r_e\left(\frac{1}{1-\alpha}\right) \quad . \quad . \quad . \quad . \quad . \quad (17)$$

when r_c is large compared with r_e, r_b and R_L. Hence, if the input signal v_s is fed from a generator resistance r_s, the input signal current i_{bs} is given by

$$i_{bs} = \frac{v_s}{r_s + r_b + r_e\left(\dfrac{1}{1-\alpha}\right)} \quad . \quad . \quad . \quad . \quad (18)$$

As in the common-base connection the variation of r_e with emitter current gives rise to cross-over distortion, which can be overcome by increasing r_s and V_{bias}; in this arrangement it is important to realize that the emitter resistance is multiplied by the factor $1/(1-\alpha)$. This factor will have values of the order of 10–50, and hence the generator resistance must be 10–50 times greater than in the common-base connection if a similar degree of linearity is to be obtained with the same forward bias voltage.

The small-signal current gain in the common-emitter connection, α_{cb}, is given by

$$\alpha_{cb} = \frac{\alpha}{1-\alpha} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (19)$$

The corresponding collector current is therefore given by

$$i_{cs} = i_{bs}\left(\frac{\alpha}{1-\alpha}\right) \quad . \quad . \quad . \quad . \quad . \quad (20)$$

Small variations in α, which is approximately unity for junction transistors, produce large variations in α_{cb}, as shown in Fig. 4.

When the input current is a linear function of input voltage, as is obtained when the generator resistance is high, this decrease in α_{cb} at high emitter currents produces distortion in a large-signal amplifier and hence the generation of harmonics. Using a push-pull circuit arrangement with matched transistors the even harmonics cancel out, leaving only the odd harmonics.

As shown in Fig. 3 a first approximation of the variation of the current gain α with emitter current is of the form

$$\alpha = \alpha_0 - \beta i_e$$

for emitter currents greater than about 1–2mA.

By the use of this formula, the large-signal collector/base-current equation becomes

$$i_{cs} = i_{bs}\left(\text{mean value of } \frac{\alpha}{1-\alpha} \text{ over the range of base current } 0 - i_{bs}\right)$$

$$\text{or } i_{cs} = i_{bs}\left(\text{mean value of } \frac{\alpha}{1-\alpha} \text{ over the range of emitter current } 0 - i_{es}\right) \quad . \quad (21)$$

Therefore

$$i_{cs} = i_{bs}\frac{1}{|i_{es}|}\int_0^{|i_{es}|}\frac{\alpha_0 - \beta i_e}{1 - \alpha_0 + \beta i_e}di_e \quad . \quad . \quad . \quad (22)$$

The modulus values of i_{es} are taken since in the push-pull arrangement considered the current gain variation is symmetrical about the zero value.

This expression for i_{cs} cannot be easily analysed to obtain the harmonic distortion for a sine-wave input signal.

A less accurate approximation is to assume that the output current waveform consists of the fundamental frequency and a given number of odd harmonics, i.e.

$$i_{cs} = A \sin \omega t + AB_1 \sin 3\omega t + \ldots + AB_n \sin (2n+1)\omega t \quad . \quad (23)$$

The input signal is a sine wave given by

$$i_{bs} = \sin \omega t \quad . \quad . \quad . \quad . \quad . \quad . \quad (24)$$

This equation has $(n+1)$ unknowns; hence the equation for i_{cs} must be solved at $(n+1)$ points on the input waveform for a given value of peak emitter current.

The simplest case to consider is that in which the output waveform consists only of the fundamental and third-harmonic frequencies:

$$i_{cs} = A \sin \omega t + AB \sin 3\omega t \quad . \quad . \quad . \quad . \quad (25)$$

There are then only two unknowns, A, the mean current gain at the fundamental frequency, and B, the fractional third-harmonic content. Hence only two equations are required.

Two possible equations are ones relating first to the tangent to the i_{cs} curve at $t = 0$, and secondly to the peak value of i_{cs}. Thus, considering the first relationship, from eqn. (25)

$$\left.\frac{di_{cs}}{dt}\right|_{t=0} = \omega(A + 3AB) \quad . \quad . \quad . \quad . \quad (26)$$

i.e. at $\quad t = 0 \quad \dfrac{di_{cs}}{di_{bs}}\dfrac{di_{bs}}{dt} = \omega(A + 3AB) \quad . \quad . \quad . \quad (27)$

but at $\quad t = 0 \quad di_{cs}/di_{bs} = \alpha_0/(1-\alpha_0) \quad . \quad . \quad . \quad (28)$

and $\quad \left.\dfrac{di_{bs}}{dt}\right|_{t=0} = \omega \cos \omega t\Big|_{t=0} = \omega \quad . \quad . \quad . \quad . \quad (29)$

therefore, from eqns. (27), (28) and (29),

$$\frac{\alpha_0}{1-\alpha_0} = A + 3AB \quad . \quad . \quad . \quad . \quad . \quad (30)$$

From eqn. (25), assuming that the peak collector current occurs when the input current is at its peak,

$$i_{c\,peak} = A - AB \quad . \quad . \quad . \quad . \quad . \quad . \quad (31)$$

but from eqn. (22)

$$i_{c\,peak} = \frac{1}{|i_{ep}|}\int_0^{|i_{ep}|}\frac{\alpha_0 - \beta i_e}{1-\alpha_0+\beta i_e}di_e \quad . \quad . \quad . \quad (32)$$

where i_{ep} is the peak emitter current.

Integrating this gives

$$i_{c\,peak} = \left\{-1 + \frac{1}{\beta i_{ep}}\log_\varepsilon\left[\frac{(1-\alpha_0)+\beta i_{ep}}{1-\alpha_0}\right]\right\} \quad . \quad (33)$$

$$= (A - AB) \text{ from eqn. (31).}$$

Hence, from eqns. (30), (31) and (33),

$$\frac{A-AB}{A+3AB} = \frac{1-B}{1+3B}$$

$$= \frac{1-\alpha_0}{\alpha_0}\left\{-1 + \frac{1}{\beta i_e}\log_\varepsilon\left[\frac{(1-\alpha_0)+\beta i_{ep}}{(1-\alpha_0)}\right]\right\} \quad . \quad (34)$$

By substitution of the known values of α_0, β and i_{ep} the fractional third-harmonic content B can be obtained. For example, from Fig. 3 $\alpha_0 = 0\cdot985$ at zero emitter current (this value would be slightly less with forward bias), $\beta = 0\cdot00077$, and for a peak emitter current of 50mA we obtain, by substitution in eqn. (34),

$$\frac{1-B}{1+3B} = 0\cdot57$$

Thus $$B = 0\cdot15$$

i.e. 15% third-harmonic distortion. With this high value of distortion the fifth harmonic would also be appreciable, but for $B < 0\cdot05$ the fifth harmonic can be neglected.

The calculation can be simplified at a small expense in accuracy if a further approximation is made. In this case it is assumed that α_{cb} has a variation of the form

$$\alpha_{cb} = \alpha_{cb0} - ki_e \quad . \quad . \quad . \quad . \quad . \quad . \quad (35)$$

Then the mean current gain A over the range of emitter current $0 - i_{e\,peak}$ is given by

$$A = \frac{\alpha_{cb0} + \alpha_{cb1}}{2}$$

where $\alpha_{cb1} = \alpha_{cb}$ at $i_{e\,peak}$, and since

$$A + 3AB = \alpha_{cb0} = \frac{\alpha_0}{1-\alpha_0}$$

B can be calculated.

For example, for the case considered this approximation gives the result $B = 0\cdot16$; i.e. this approximation gives a slightly higher value of distortion.

In both approximations it is necessary to know the value of i_{ep}. The peak collector current is known, since R_L and V_{ht} are fixed. For the purposes of calculation it is sufficiently accurate to estimate α_m, the mean value of α from Fig. 3, in which case

$$i_{ep} = \frac{i_{c\,peak}}{\alpha_m} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (36)$$

Using this second approximation the following results are obtained: For less than 5% third-harmonic distortion α_{cb1} must be greater than $0\cdot74\alpha_{cb0}$; for less than 10% distortion α_{cb1} must be greater than $0\cdot54\alpha_{cb0}$; and for less than 15% distortion α_{cb1} must be greater than $0\cdot38\alpha_{cb0}$.

The current gain A at the fundamental frequency can be obtained from eqn. (30), which gives

$$A(1+3B) = \alpha_0/(1-\alpha_0)$$

and hence $$A = \frac{\alpha_0}{(1-\alpha_0)(1+3B)} \quad . \quad . \quad . \quad . \quad . \quad (37)$$

The a.c. power developed in the load resistor R_L at the fundamental frequency is therefore given by the expression

$$P_{load} = A^2 i_{bs}^2 R_L$$

Hence $$P_{load} = \left(\frac{\alpha_0}{1-\alpha_0}\right)^2\left(\frac{1}{1+3B}\right)^2\frac{v_s^2 R_L}{\left[r_s + r_b + \left(\frac{1}{1-\alpha}\right)r_e\right]^2}$$

by substitution from eqns. (37) and (18).

This becomes

$$P_{load} \simeq \left(\frac{\alpha_0}{1-\alpha_0}\right)^2\left(\frac{1}{1+3B}\right)^2\frac{R_L v_s^2}{r_s^2} \quad . \quad . \quad . \quad (38)$$

since r_s is large compared with $r_b + \left(\frac{1}{1-\alpha}\right)r_e$.

Therefore the power gain of the common emitter amplifier is given by

$$P_{gain} = \left(\frac{\alpha_0}{1-\alpha_0}\right)^2\left(\frac{1}{1+3B}\right)^2\frac{4R_L}{r_s} \quad . \quad . \quad . \quad (39)$$

For the transistor shown in Fig. 3, when $r_s = 5$ kilohms, $R_L = 400$ ohms, and $V_{ht} = 20$ volts, the peak current = 50mA, the a.c. power output is approximately 500mW, $B = 0\cdot15$ and the power gain of the fundamental frequency is approximately 400.

(6.3) Common-Collector Amplifier

The common-collector-circuit arrangement is shown in Fig. 9. The input signal is applied via the phase-splitting transformer T_1

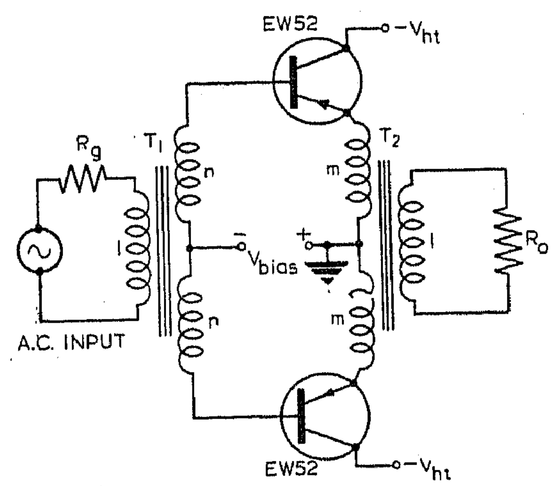


Fig. 9.—Basic circuit of the common-collector push-pull amplifier.
$r_s = n^2 R_g$
$R_L = m^2 R_0$

to the base electrodes, and the load transformer T_2 is connected to the emitter electrodes. The base electrodes are biased forward by the negative voltage V_{bias} applied to the centre tap of the secondary winding of transformer T_1.

The input resistance of a transistor in the common-collector connection at low frequencies is given by the expression[4]

$$r_{in} = r_b + \frac{r_c}{1 + \dfrac{(1-\alpha)(r_c + r_b)}{R_L + r_e}}$$

or
$$r_{in} = r_b + \left(\frac{1}{1-\alpha}\right)(r_e + R_L) \quad . \quad . \quad . \quad (40)$$

when R_L is small compared with $r_c(1-\alpha)$, as is generally the case.

Hence, if the input signal v_s is fed from a generator of resistance r_s, the input signal current i_{bs} is given by

$$i_{bs} = \frac{v_s}{r_s + r_b + \left(\dfrac{1}{1-\alpha}\right)(r_e + R_L)} \quad . \quad . \quad . \quad (41)$$

and the voltage developed across the load resistor R_L is given by

$$v_L = i_{es}R_L = R_L \times i_{bs} \times \alpha_{eb} \quad . \quad . \quad . \quad (42)$$

where α_{eb} is the current gain between the emitter and base electrodes. This is related to the current gain α by the expression

$$\alpha_{eb} = \frac{1}{1-\alpha} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (43)$$

Therefore
$$v_L = R_L \times i_{bs} \times \frac{1}{1-\alpha}$$

Hence
$$v_L = R_L \times \frac{1}{1-\alpha} \times \frac{v_s}{r_s + r_b + \left(\dfrac{1}{1-\alpha}\right)(r_e + R_L)} \quad . \quad (44)$$

As in the other circuit arrangements, the variation in emitter resistance produces cross-over distortion, but in the common-collector arrangement the emitter is in series with the load resistor and therefore this cross-over distortion can be eliminated by forward bias alone.

From eqn. (44) the voltage gain of the common-collector amplifier is given by

$$\left.\begin{aligned} \frac{v_L}{v_s} &\simeq \frac{\left(\dfrac{1}{1-\alpha}\right)R_L}{r_s + \left(\dfrac{1}{1-\alpha}\right)R_L} \\ \text{or} \quad \frac{v_L}{v_s} &\simeq \frac{1}{\dfrac{r_s(1-\alpha)}{R_L} + 1} \end{aligned}\right\} \quad . \quad . \quad . \quad (45)$$

where R_L is large compared with r_e and the base resistance r_b is included in r_s.

Hence, substituting the variation in α as given by the expression

$$\alpha = \alpha_0 - \beta i_e$$

we obtain
$$\frac{v_L}{v_s} = \frac{1}{\dfrac{r_s}{R_L}(1-\alpha_0) + 1 + \dfrac{r_s}{R_L|i_{es}|}\displaystyle\int_0^{|i_{es}|}\beta i_e di_e} \quad . \quad . \quad (46)$$

where i_{es} is the instantaneous value of i_e corresponding to the input voltage v_s.

Therefore
$$\frac{v_L}{v_s} = \frac{1}{\dfrac{r_s}{R_L}(1-\alpha_0) + 1 + \dfrac{\beta r_s i_{es}}{2R_L}} \quad . \quad . \quad . \quad (47)$$

This equation cannot easily be analysed for a sine-wave input voltage, but by assuming a given number of harmonics an approximate solution can be obtained.

The analysis of this amplifier arrangement is similar to that of the common-emitter amplifier, except that this time it is more convenient to consider voltage gains. Thus we apply a sine wave input voltage $V_{bs} = \sin \omega t$ and consider the output voltage to be of the form

$$V_{es} = A \sin \omega t + AB \sin 3\omega t$$

where A is the mean voltage gain at the fundamental frequency and B is the fractional third-harmonic component.

Then, using eqn. (47), we find the voltage gain in the quiescent state to be given by

$$\frac{1}{\dfrac{r_s}{R_L}(1-\alpha_0) + 1}$$

But this is also given by

$$\left.\frac{dV_{cs}}{dV_{bs}}\right|_{t=0} = \left.\frac{dV_{es}}{dt}\frac{dt}{dV_{bs}}\right|_{t=0} = A + 3AB$$

so that
$$A + 3AB = \frac{1}{\dfrac{r_s(1-\alpha_0)}{R_L} + 1} \quad . \quad . \quad . \quad . \quad (48)$$

and, assuming that the peak output voltage occurs at the same time as the peak input voltage, we obtain the relationship

$$A - AB = \text{(peak output voltage)} = \frac{1}{\dfrac{r_s}{R_L}(1-\alpha_0) + 1 + \dfrac{\beta r_s i_{ep}}{2R_L}} \quad . \quad . \quad . \quad (49)$$

Therefore

$$\frac{A + 3AB}{A - AB} = \frac{1 + 3B}{1 - B} = \frac{\dfrac{r_s}{R_L}(1-\alpha_0) + 1 + \dfrac{\beta r_s i_{ep}}{2R_L}}{\dfrac{r_s}{R_L}(1-\alpha_0) + 1} \quad . \quad (50)$$

The values of β and α_0 are inherent in the transistor, and i_{ep} and R_L are fixed by the output-power requirements and h.t. voltage, but by varying the generator resistance, r_s, the amount of third-harmonic distortion can be varied. In effect, decreasing the generator resistance increases the negative feedback in the transistor circuit and reduces distortion.

Fig. 10 gives the percentage distortion as a function of r_s (calculated for the transistor used in obtaining Fig. 3) for two values of peak emitter current, showing the reduction in harmonic distortion with decreasing r_s.

In the common-collector circuit the voltage gain A of the fundamental component of the input signal is obtained from eqn. (48) and is given approximately by

$$A = \frac{R_L}{r_s(1-\alpha_0) + R_L}\frac{1}{1+3B} \quad . \quad . \quad . \quad (51)$$

Therefore the a.c. power into the load resistance is

$$P_{load} = \frac{A^2v_s^2}{R_L} = \frac{R_L v_s^2}{[r_s(1-\alpha_0) + R_L]^2}\frac{1}{(1+3B)^2} \quad . \quad (52)$$

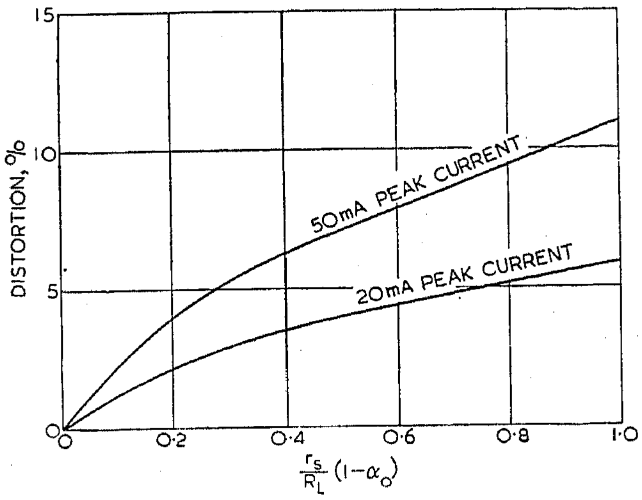


Fig. 10.—Theoretical percentage third-harmonic distortion for a common-collector output stage as a function of generator impedance r_s.

The power gain is therefore given by

$$P_{gain} = \frac{4R_L r_s}{[r_s(1-\alpha_0)+R_L]^2(1+3B)^2} \quad . \quad . \quad (53)$$

With the same h.t. voltage and power-output requirement as in the common-emitter amplifier example considered, and using a generator of resistance 4 kilohms, the power gain of the common-collector amplifier is approximately 25; this is less than in the common-emitter connection, but the third-harmonic distortion has been reduced by a factor of three (4·5% compared with 15%).

The a.c. power required to drive a transistor amplifier cannot be obtained direct from the output power required and the power gain of the stage as defined by the expression

$$P_{gain} = \frac{\text{A.C. power into load}}{\text{Maximum a.c. power available from the input generator}}$$

since in all three circuit arrangements the generator is not matched to the input resistance.

All transistor amplifiers, unlike those using thermionic valves, require an appreciable amount of input power. Hence, if the driver stage is operated inefficiently, the output-power requirement of the preceding amplifier becomes greater. In considering the overall performance of an output stage it is important to know the total amount of power developed in the output circuit of the driver stage (including its load).

Describing the output of the driver stage as a generator of internal resistance r_s and open-circuit voltage v_s, the total driver-stage power developed, P_{gen}, when operating into a load R_{in}, is given by

$$P_{gen} = \frac{v_s^2}{r_s + R_{in}} \quad . \quad . \quad . \quad . \quad . \quad . \quad (54)$$

In general, R_{in} is the mean value of r_{in}, the input resistance of the output stage, over the range of input signal v_s.

A useful method of comparing and assessing the performance of the three basic output-stage arrangements is by means of the factor G, defined by the expression

$$G = \frac{P_{load}}{P_{gen}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (55)$$

where P_{load} is the power developed in the load of the output stage and P_{gen} is the total power developed in the driver stage as defined above.

In the discussion on the common-base and common-emitter arrangements, it was shown that R_{in} is generally small compared with r_s and hence

$$P_{gen} \simeq \frac{v_s^2}{r_s}$$

giving

$$G \simeq \frac{P_{load} r_s}{v_s^2}$$

$$= \frac{P_{gain}}{4}$$

where P_{gain} is equal to $\dfrac{P_{load}}{v_s^2/4r_s}$.

In the common-collector arrangement, on the other hand, the input resistance R_{in} is generally greater than r_s. From eqn. (40) the small-signal input resistance, r_{in}, is given by

$$r_{in} \simeq r_b + \frac{R_L}{1-\alpha}$$

and hence

$$R_{in} \simeq r_b + \frac{R_L}{1-\alpha_m}$$

where α_m is the mean value of α over the emitter-current range zero–i_{ep}.

Thus

$$P_{gen} = \frac{v_s^2}{r_s + r_b + R_L/(1-\alpha_m)}$$

and

$$G = \frac{P_{load}}{v_s^2}\left(r_s + r_b + \frac{R_L}{1-\alpha_m}\right)$$

which is greater than $\dfrac{P_{load} r_s}{v_s^2}$

From the performance of the three arrangements when delivering an output of 500 mW and operating from an h.t. voltage of 20 volts, the results shown in Table 1 are obtained.

Table 1

Comparison of the Three Amplifier Arrangements

Arrangement	r_s	P_{gain}	G
		dB	dB
Common base	100	12	6
Common emitter	5 000	26	20
Common collector . .	4 000	14	14

During the non-conducting half-cycle in the common-base and common-emitter arrangements, the emitter electrode is cut off and a peak collector voltage equal to twice the h.t. voltage is reflected via the output transformer.[5] In the common-collector amplifier the collector is kept at the h.t. voltage, but during the cut-off half-cycle a peak reverse voltage approximately equal to the h.t. voltage is applied to the base; hence the peak collector/base voltage is equal to twice the h.t. voltage. Thus in all three circuit arrangements the h.t. voltage must be less than one-half the maximum collector voltage, $V_{c\,max}$.

Two EW52 transistors matched for variations in current gain with emitter current were connected in a common-collector-circuit arrangement with the d.c. bias conditions as detailed in Table 2. Under these conditions the maximum a.c. output power into the output transformer was 100 mW. The total input power (both d.c. and a.c.) was 146 mW, giving a power efficiency of 69%. The difference between this value and the theoretical

Table 2

Current gain α_0 (at $i_c = 0{\cdot}6$mA)	0·9775
Variation in gain β	0·000 97/mA
H.T. voltage	10·5 volts
Bias voltage	150mV
Total standing collector current	1·2mA (0·6mA per transistor)
Output Transformer. (T_2)	
Turns ratio	12·5 + 12·5 : 1
Primary d.c. resistance	7 ohms (each half)
Secondary d.c. resistance	0·2 ohm
Theoretical primary reflected impedance of 3-ohm load	470 ohms
Practical primary impedance (at 1 kc/s) ..	530 ohms
Input Transformer. (T_1)	
Turns ratio	1 : 1·5 + 1·5
Primary input impedance	6 kilohms
Experimental Results	
Maximum r.m.s. voltage across 3-ohm load resistor (without peak clipping)	0·52 volt
Maximum power into 3-ohm load	90mW
Maximum power into output transformer (T_2)	100mW
Under these conditions, mean collector current	13·5mA
Power drain from h.t. supply	142mW
Input signal power	4mW
Total input power	146mW
Power efficiency (neglecting transformer loss)	69%
Power drain from h.t. supply during quiescent state	13mW
Total input power less quiescent power drain	(146 − 13) = 133mW
Using this value of input power gives an efficiency of	76%

maximum value of 78% was almost entirely due to a dissipation of 13mW of power in the quiescent state.

Excluding this quiescent-state power drain gives a power efficiency of 76%.

With the output power of 100mW, the mean power dissipated in each transistor was 22·5mW. When the h.t. voltage was increased to 22·5 volts an output power of 460mW was obtained with a power dissipation in each transistor of 105mW.

Using an h.t. voltage of 10·5 volts, the harmonic distortion of the output signal was measured for two values of peak current and for various values of generator resistance.

The results obtained for third-harmonic distortion are shown in Fig. 11, together with the theoretical values calculated from eqn. (50). These are seen to be in reasonably good agreement.

The fifth-harmonic distortion was found to be negligible, having a value less than 0·2%, while the second-harmonic content had values between 0·5% and 1·5%, the higher values occurring when the generator resistance was high.

One practical consideration favouring the choice of the common-collector arrangement is that the design of the output transformer T_2 can be simpler than in the other arrangements. In the common-base and common-emitter arrangements the output resistance of the amplifiers is considerably higher than the load resistance R_L. Hence the low-frequency response of the amplifier is determined by the time-constant of R_L and the primary inductance of transformer T_2. In the common-collector circuit the output resistance R_{out} [approximately $r_s(1 - \alpha_0)$] of the transistor for low distortion is less than R_L, and thus the low-frequency response is now determined by the time-constant of the effective resistance, R_L and R_{out} in parallel and the primary

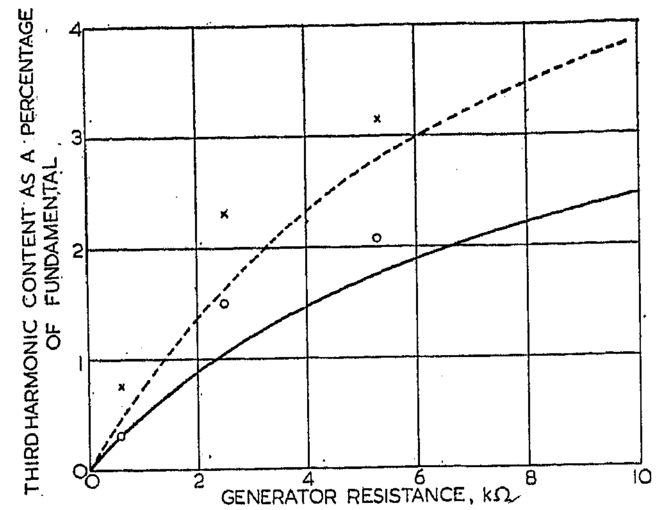


Fig. 11.—Measured third-harmonic distortion of a common-collector output stage as a function of generator resistance.

——— Calculated, ○○○ Experimental } $i_{ep} = 9{\cdot}7$mA.
– – – – Calculated, ××× Experimental } $i_{ep} = 15{\cdot}3$mA.

inductance of T_2. For the same value of primary inductance this time-constant is therefore lowest for the common-collector arrangement, which will thus give a better performance at low frequencies. Alternatively for the same low-frequency response as the other amplifier arrangements it is possible to use a lower value of primary inductance in the common-collector case. Similarly, the effects of shunt capacitances due to the transformer at high frequencies are also less when using the common-collector arrangement.

(7) THE DRIVER STAGE FOR THE COMMON-COLLECTOR OUTPUT STAGE

Since the common-collector arrangement offers a useful compromise between gain and low distortion, its design has been considered in the greatest detail. One of the main considerations is the design of the driver stage.

The common-collector output stage has been shown to have a relatively high input resistance but requires a large driving voltage, the peak-to-peak voltage being twice the h.t. voltage.

Thus to obtain the necessary balanced drive, and using the same h.t. voltage supply as for the output stage, a phase-splitting transformer is necessary. For a half-watt output stage, the power required from the driver stage is about 10–20mW.

The primary resistance ($R_{primary}$) of the phase-splitting transformer is determined by the collector voltage and standing primary current, since the driver stage is a large-signal power amplifier and cannot be correctly matched. The relationships using a Class A driver amplifier are

$$\left.\begin{aligned} &\text{Power required} < \frac{V_c^2}{2R_{primary}} \\ \text{and}\quad &\text{Power required} < \frac{i_t^2 R_{primary}}{2} \end{aligned}\right\} \quad . \quad . \quad (56)$$

where V_c is the driver-stage collector voltage and i_t is the standing current through the transformer primary. These relationships merely set the limits to prevent "clipping" of the sine-wave output signal from the driver stage. In order to avoid introducing distortion at this stage, V_c and i_t are made considerably greater than the maximum values called for by the inequalities given in expression (56).

A typical example is an audio power of 20mW with a collector voltage of 5 volts; then $i_t = 10$mA and $R_{primary} = 500$ ohms.

The primary winding of the phase-splitting transformer, T_1, can be connected in either the collector or the emitter leads of

the driving transistor. If a common-emitter-connected driver stage is used, the reflected load resistance (500 ohms) is fed from the collector output resistance, which will be of the order of 20–30 kilohms. This means that the source resistance presented by the stage to the output stage via the phase-splitting transformer is very high, and the driver becomes practically a constant-current source, which has been shown to be the worst possible condition for output linearity in the common-collector output stage. If a common-collector-connected driver stage is used, the reflected load resistance (which in the case considered is 500 ohms) is fed from the emitter output resistance, which is given approximately by

$$R_{out} = (1 - \alpha)(r_b + R_S) + r_e$$
$$\simeq (1 - \alpha)R_S \text{ if } R_S \gg r_b \text{ and } r_c \quad . \quad . \quad . \quad (57)$$

where R_S is the source impedance at the base electrode of the driver stage. This can have a low value compared with the load resistance (500 ohms), and is the condition for minimum distortion.

A suitable driver-circuit arrangement is shown in Fig. 12.

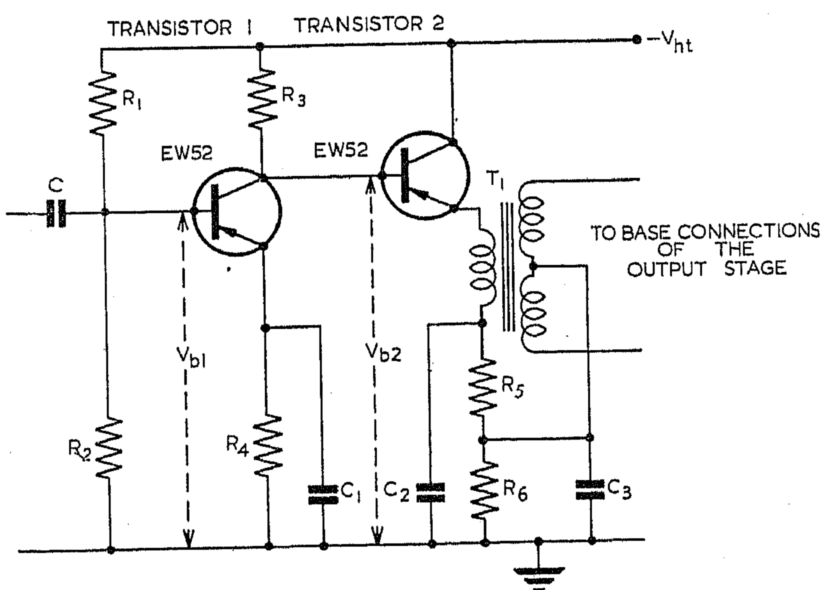


Fig. 12.—Common-collector driver stage.

The common-collector driver stage (Transistor No. 2) is fed direct from the collector of the last audio-amplifier stage (Transistor No. 1). In this circuit the d.c. emitter-current bias, and hence the d.c. collector current, I_{c1}, in the first transistor is determined by the base voltage, V_{b1} (produced by the voltage divider R_1 and R_2), which appears across the emitter resistor R_4 (in series with the d.c. emitter resistance). This collector current is virtually independent of the current gain, α_{cb}, of this stage. The collector voltage, V_{b2}, is therefore approximately $V_{ht} - R_3I_{c1}$, which appears across R_5 in series with R_6 and determines the steady current flowing through the transformer primary.

In general the collector resistance, R_3, will be much smaller than the output resistance of the collector of transistor No. 1, e.g. 3 kilohms compared with 30 kilohms. Thus the source resistance at the base of the driver stage is approximately R_3, which from eqn. (57) gives a driver output resistance of approximately

$$R_3(1 - \alpha)$$

For example, if

$$R_3 = 3 \text{ kilohms and } (1 - \alpha) = \tfrac{1}{30}$$
$$R_{out} = 3 \text{ kilohms} \times \tfrac{1}{30} = 100 \text{ ohms}$$

This output resistance is low compared with the reflected resistance of transformer T_1, and so satisfies the condition for low distortion due to variations in α of the output stage. In effect, the variable reflected resistance is shunted by a smaller linear resistance and linearity is obtained at the expense of power gain. The advantage of this circuit is that the power wasted in the shunt resistor is at a low level; e.g. for a $\frac{1}{2}$ watt output from the final output stage, the power dissipated in resistor R_3 is approximately 1–2 mW.

The standing current through the driver transistor is also used to produce the low-impedance forward bias to the output circuit, using resistor R_6.

The turns ratio of the step-up transformer T_1 is determined by $R_{primary}$ [expression (56)] and the input resistance of the final stage. Thus

$$n \simeq \sqrt{\left[\frac{R_L}{(1 - \alpha_0)} \frac{1}{R_{primary}}\right]}$$

(8) COMPLEMENTARY SYMMETRY

Sziklai[3] has described interesting push-pull connections using a *p–n–p* and *n–p–n* transistor in a complementary-symmetry circuit arrangement. These have the advantage that the input is then single-ended, which eliminates the phase-splitting transformer (T_1) necessary in the conventional circuits previously described. However, both positive and negative h.t. voltages are necessary.

The input circuit of a transistor under Class B conditions is in effect a diode, the input resistance being high when the transistor is non-conducting and low when it is conducting; thus if each half of the complementary-symmetry circuit is fed from separate capacitors these will charge up so that only the peaks of the input signal will be amplified. An alternative method is to feed the two inputs from a single capacitor as shown in Fig. 13; for a common-collector arrangement this

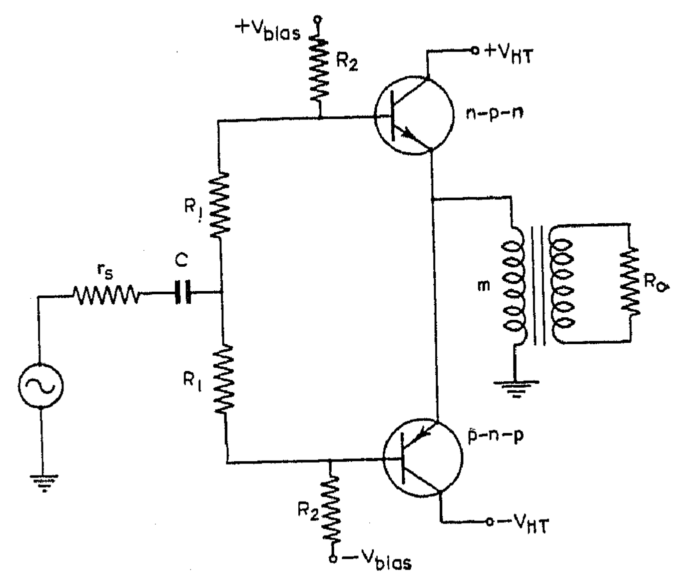


Fig. 13.—Complementary-symmetry type of common-collector amplifier using a single input capacitor.

eliminates the rectifying action, provided that the input impedances of the two transistors are matched at all values of input signal.

To reduce cross-over distortion, forward bias must be applied to the input electrode, which in the circuit shown in Fig. 13 is the base electrode. For the *p–n–p* transistor the base must be biased negatively, and for the *n–p–n* transistor the base must be biased positively; therefore the resistors R_1 (Fig. 13) must be connected in series with the base electrodes to drop the bias voltages that are applied via resistors R_2. Resistances R_2 must be large compared with the input impedance of the transistors, $[\alpha_0/(1 - \alpha_0)]R_L$, to avoid shunting the input signal.

As shown in Section 6, the common-collector amplifier gives a good compromise between power gain and distortion. Using complementary symmetry it has the further advantage that a single-ended output as well as input can be used (Fig. 13).

However, the peak-to-peak drive voltage is equal to twice the h.t. voltage, so that with capacitive coupling the driver stage must have an h.t. voltage greater than twice that of the output stage.

A serious shortcoming of complementary-symmetry-type large-signal amplifiers is that the variation of current gain α with emitter current differs in *p–n–p* and *n–p–n* transistors of equal emitter area. This gives rise to even-harmonic distortion. It is therefore necessary that the emitter of the *p–n–p* transistor should be larger in area by a factor of approximately 2 than that in the *n–p–n* transistor.

(9) NON-SINUSOIDAL APPLICATIONS

In many applications the waveform being amplified is not sinusoidal and distortion is unimportant. Examples are, first, the use of a transistor as an on-off switch, the operation being similar to that of a relay, and secondly, circuits used for converting low direct voltages to high direct voltages.

In these applications the transistor is generally required to be either fully "on" or "off." Usually the two states are defined as follows: when the transistor is "off," the emitter current is zero and the collector is approximately at the h.t. potential, the collector current being I_{c0}, which is generally less than $0 \cdot 1$ mA. In the "on" condition the input current is such that the maximum possible current flows through the collector, i.e. almost all the h.t. voltage is dropped across the collector load resistance, the voltage appearing at the collector being usually less than $0 \cdot 1$ volt. Thus in either condition the actual dissipation in the collector itself is very low. At any intermediate condition it is possible to dissipate large powers in the collector, and thus it is essential when switching from one condition to the other to do so in a time which is short compared with the thermal time-constant τ_T of the transistor, otherwise the maximum power that can be switched must be considerably reduced. To calculate the mean dissipation in the transistor it is necessary to know:

(*a*) The I_{c0} flowing when $V_c = V_{ht}$.

(*b*) The voltage dropped across the transistor when it is bottomed, i.e. in the "on" state.

(*c*) The rate at which the transistor is switched from "off" to "on" and vice versa.

(*d*) The duration of the "on" and "off" states.

The limitation to the power that can be produced in R_L, the load resistance, by switching from "off" to "on" is entirely governed by P_{cmax}, the mean power dissipation limit set for the transistor. When the transition from "off" to "on" is rapid compared with τ_T, powers of the order of $10P_{cmax}$ can be developed in R_L.

Thus in a practical case where a single low-power junction transistor was made to operate a relay, the mean power developed in the relay coil was 450mW, whilst the mean power dissipated in the transistor was 45mW. Since the operation was on a 1 : 1 "on" to "off" basis, the peak power in the relay (i.e. the power during the "on" period) was 900mW. In this particular case the time taken to switch from "off" to "on" was 2 millisec.

(10) CONCLUSIONS

Transistors can be used to advantage in power-amplifier circuits because of their high efficiency. It has been shown that the maximum output that can be handled, in sinusoidal applications, is not necessarily limited by the maximum collector dissipation rating. Thus transistors for these applications must be so designed that the variation of the current gain over the required range of emitter current be kept below a value determined from the amount of distortion that can be tolerated. If the maximum collector voltage could be increased by improved design, the peak emitter current would be lower and hence the important of the current gain variation would be reduced.

The cross-over distortion in Class B amplifiers due to the variation of emitter resistance with emitter current can be reduced by simple circuit techniques.

In view of the power limitations of existing junction transistors, push-pull Class B operation is considered to be more important than Class A. Of the three possible basic amplifier arrangements the common-collector type is preferred because it provides a useful compromise between gain and distortion and also because the design of the output transformer is simpler.

Complementary symmetry, although it is in some ways attractive, does not offer many advantages in the type of amplifier applications considered in the paper.

In certain non-sinusoidal applications the mean power dissipated in the collector sets the limit to the amount of power that can be handled. Under certain conditions the mean output power can be many times the mean power dissipated in the transistor—which can therefore be used, for example, as a very efficient current switch.

(11) REFERENCES

(1) SABY, J. S.: "Transistors for High Power Application," Convention Record of the Institute of Radio Engineers 1954, Part 3, p. 80.

(2) GIACOLETTO, L. J.: "Power Transistors for Audio Circuits," *Electronics*, January, 1954, p. 144.

(3) SZIKLAI, G. C.: "Symmetrical Properties of Transistors and their Applications," *Proceedings of the Institute of Radio Engineers*, 1953, **41**, p. 717.

(4) SHEA, R. F. (Editor): "Principles of Transistor Circuits" (Wiley, New York, 1953).

(5) TERMAN, F. E.: "Radio Engineers Handbook" (McGraw-Hill Book Co., London, 1950).

(6) MISSEN, J. I.: "A Crystal Triode Push-Pull Amplifier," *G.E.C. Journal*, 1953, **20**, No. 3, July.

(7) WEBSTER, W. M.: "On the Variation of Junction Transistor Current Amplification Factor with Emitter Current," *Proceedings of the Institute of Radio Engineers*, 1954, **42**, p. 914.

[The discussion on the above paper will be found on page 786.]

621.314.5 : 621.373.42 : 621.314.7

Paper No. 1862 R
Apr. 1955

TRANSISTOR D.C. CONVERTORS

By L. H. LIGHT, B.Sc., and PRUDENCE M. HOOKER, B.Sc.

(*The paper was first received* 3*rd March, and in revised form* 18*th March*, 1955. *It was published in April*, 1955, *and was read before a* JOINT MEETING *of the* RADIO *and* MEASUREMENTS SECTIONS 11*th May*, 1955.)

SUMMARY

A transistor relaxation-oscillator circuit serving as a direct-voltage step-up device is described. The transistor acts as an interrupter in a low-voltage circuit; energy is stored during the "on" period in the transformer inductance and is delivered to the output circuit at an increased voltage during the "off" period. The paper contains a full account of its operating principles and design for efficient power conversion, a brief analysis of voltage-doubler variants and a discussion of the performance and applications of practical d.c. convertors. It also presents some methods of dealing with problems associated with the circuit, such as stabilization of output voltage and fault protection.

Transistor convertors have several advantages over competitive methods at power levels of a fraction of a milliwatt up to several watts, such as high efficiency, long life and small size. They therefore promise to become the preferred method of generating h.t. voltages from l.t. supplies at low power levels, except when temperature conditions are unsuitable for the semi-conductor elements employed.

LIST OF SYMBOLS

A = Area of cross-section of transformer core.
B_{pk} = Peak operating flux density.
f = Operating frequency.
i_b = Base current.
I_{bb} = Base current during input stroke.
i_c = Collector current.
I_{c0} = Collector cut-off current.
I_i = Mean input current.
I_0 = Output current.
i_{pk} = Peak input current in primary winding.
i_{spk} = Peak current in secondary winding.
i_s = Current in secondary winding.
k = Optimum primary/secondary winding-space ratio.
l = Flux path length in core.
l_a = Flux path length in air-gap.
L_p = Inductance of primary winding.
L_s = Inductance of secondary winding.
N_p = Number of turns in primary winding.
N_b = Number of turns in base winding.
N_s = Number of turns in secondary winding.
P_0 = Output power.
r_b = D.C. base input resistance of transistor.
R_b = External resistance in base circuit.
r'_c = Collector resistance of bottomed transistor.
R_i = Internal resistance of battery.
R_L = Load resistance.
R_0 = Output impedance.
R_p = Resistance of primary winding.
t_f = Duration of input stroke.
t_b = Duration of output stroke.
T = Period of cycle = $(t_f + t_b)$.
V_{bb} = Transistor base input voltage required to give $i_c = i_{pk}$.
V_{bw} = Voltage developed across base winding during input stroke.
$V_{c\,max}$ = Maximum collector—base voltage rating of transistor.
V_{df} = Forward-voltage drop in rectifier D_1.
V_i = Input voltage.
V_0 = Output voltage.
V_q = Collector voltage during output stroke.
V_Q = Maximum permissible collector voltage during output stroke.
V_r = Total negative collector—base voltage during output stroke.
V_{spk} = Peak voltage across secondary winding.
α' = Static base—collector current gain of transistor.
η = Efficiency.
μ = Permeability.

Mr. Light and Mrs. Hooker are with the Mullard Radio Valve Co., Ltd.

(1) INTRODUCTION

Of the many methods which are used to convert d.c. power from one voltage to another, all except the rotary convertor use some type of interrupter in the input circuit. This may take the form of a mechanical vibrator (as in the conventional vibrator power supply), a thermionic valve (as in the radio-frequency or ringing-choke supply), or a relaxation-oscillator circuit using either a thyratron or a cold-cathode tube.

Soon after the introduction of transistors it was appreciated that they also could be used as efficient low-voltage interrupters by using the property of "bottoming"—a condition of current saturation in which transistors can pass large currents with only a small voltage drop across them. Early workers successfully used point-contact transistors in voltage convertors in a variety of oscillator circuits.[1,2] The junction transistor, which has since become available, has particularly favourable characteristics for this application by virtue of its very low resistance in the "on" state and its high resistance in the "off" state.[3]

The paper presents an account of a study of d.c. convertors using as their basis a particular junction transistor relaxation-oscillator circuit invented by P. H. J. Janssen and C. v. d. Vijver. While it is also possible to make successful d.c. convertors with other circuits,[4,5] such as other relaxation or class C oscillators, the circuit described has the advantages of simplicity, high conversion efficiency, and of being non-critical in design and readily adjusted to allow for transistor spreads.

(2) OPERATION OF CIRCUIT

(2.1) Operating Principles

The circuit (see Fig. 1) is a simple transformer-coupled relaxation oscillator in which energy is stored in the inductance of the transformer during the "on" period of the transistor (the input stroke) and is delivered to the output circuit at a higher voltage during the "off" period (the output stroke). During the input stroke the transistor is bottomed and practically the whole of the input voltage is applied across the transformer primary winding. A rising current flows through the transistor and primary winding, while a constant control current, I_{bb}, flows into the base of the transistor. The rectifier D_1 is non-conducting during this period. The transistor remains bottomed so long as the collector current is less than $\alpha' I_{bb}$, where α' is the static base-collector current gain of the transistor. As soon as this condition

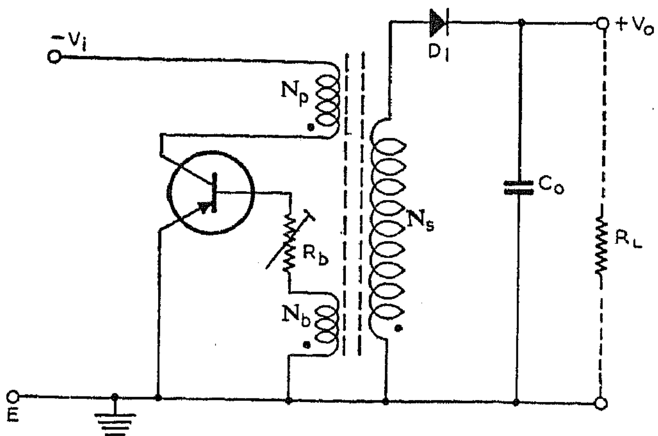


Fig. 1.—D.C. convertor circuit.

no longer holds, cumulative switch-off of the transistor occurs. The voltage across the windings then reverses and rises rapidly as ringing commences. The rise in voltage is, however, arrested as soon as the secondary voltage reaches the value of the output voltage. A current decreasing from an initial maximum then flows in the rectifier D_1, and transfers the energy stored in the inductance into the output capacitor. When this output current has fallen to zero, the input is switched on again.

It will be seen that this circuit has much in common with the system of e.h.t. generation commonly employed in television technique. A very important difference in operation, however, is that the sharp rise in transformer voltage is arrested long before the peak voltage, which would be generated during ringing, is reached.

The voltage waveforms (Fig. 3) in the circuit are essentially rectangular; the input and output voltages are the prime determining factors for conditions during the input and output strokes, respectively. The flux in the transformer does not reverse; its waveform is approximately triangular, with the maximum occurring at the end of the input stroke and the minimum (the residual induction) at the end of the output stroke.

As the flow of energy to the output alternates with the input from the battery, the loading on the output has little influence on the input stroke. The power flowing into the convertor is thus practically constant (it depends on the setting of the variable resistor R_b), and as the conversion efficiency also does not vary much with output loading, the power delivered by the convertor to the load is nearly constant irrespective of the load. This constant-power characteristic is of value in some applications; methods for deriving a constant-voltage characteristic, which is preferable in other applications, are discussed in Section 5.

(2.2) Detailed Analysis of Operation

The following approximations are made:

(*a*) The collector characteristic of the bottomed transistor (OP in Fig. 2) is assumed to be that of a low resistance r'_c.

(*b*) The transformer leakage inductance is neglected so that

$$L_s = \left(\frac{N_s}{N_p}\right)^2 L_p$$

(*c*) The loading on the transformer during the input stroke (which consists of the power dissipated in the base circuit and reverse leakage in the rectifier) is neglected.

(*d*) The resistance of the base and secondary windings is neglected.

(*e*) The d.c. base input resistance, r_b, of the bottomed transistor is assumed to be constant independent of the collector current flowing (it may actually increase somewhat with collector current).

These approximations introduce but little error into the analysis.

Absolute values of currents and voltages are used in all the expressions, but for clarity the polarities are mentioned in the

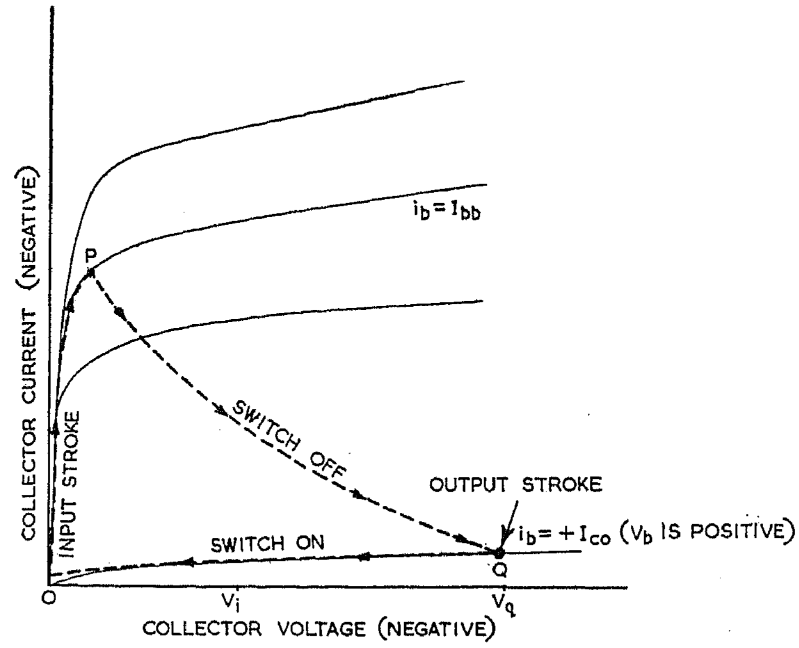


Fig. 2.—Locus of transistor operating point.

text. It is assumed throughout the paper that *p-n-p* transistors are used, and the polarities quoted are appropriate to them. *n-p-n* transistors, however, function equally well, provided that the battery and rectifier polarities are suitably changed.

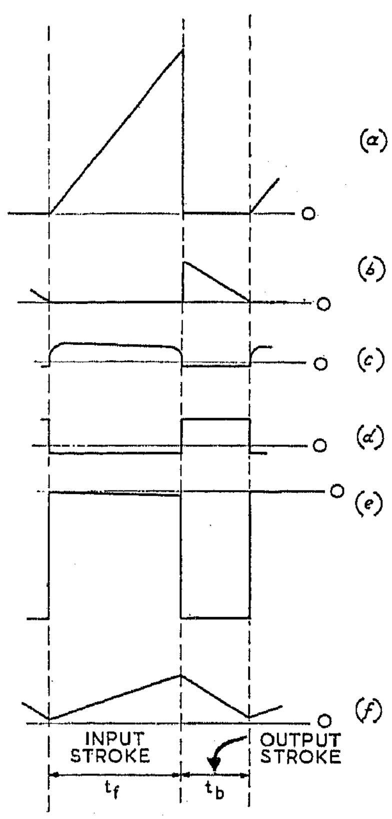


Fig. 3.—Waveforms in circuit of Fig. 1.

(*a*) Current in primary winding.
(*b*) Current in secondary winding.
(*c*) Base current (negative).
(*d*) Base voltage.
(*e*) Collector voltage.
(*f*) Flux in transformer core.

(2.2.1) Input Stroke.

The current rise in the transformer primary winding is defined by the equation

$$L_p\frac{di}{dt} + (r'_c + R_p)i = V_i \quad . \quad . \quad . \quad . \quad . \quad (1)$$

with the solution

$$i = \frac{V_i}{r'_c + R_p}[1 - \varepsilon^{-t(r'_c+R_p)/L}] \quad . \quad . \quad . \quad . \quad (2)$$

where i = Instantaneous input current.
R_p = Resistance of primary winding.
V_i = Input voltage.

Only the initial part of the rise is used, so that it is usually permissible to ignore the voltage drop across R_p and r'_c. The current can then be considered to rise linearly from a value very near to zero (which depends on the loading on the transformer) according to the equation

$$i \simeq \frac{V_i}{L_p}t \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

A voltage approximately equal to V_i therefore exists across the primary winding during the input stroke. This induces a negative voltage, V_{bw}, in the base winding given by

$$V_{bw} \simeq \frac{N_b}{N_p}V_i \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

where N_p and N_b are the number of turns in the primary and base windings, respectively. This causes an approximately constant base current, I_{bb}, to flow in the transistor, given by

$$I_{bb} \simeq \frac{V_{bw}}{R_b + r_b} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

where R_b is the resistance in series with the base winding.

The collector current rises according to eqn. (2) until it reaches the value $\alpha' I_{bb}$, when the transistor comes out of the bottoming condition. Cumulative switch-off then takes place. The collector voltage rises and the voltage across the primary winding falls, which produces a fall in the base current. This leads to a reduction in collector current which in turn causes the voltage across the windings, and in particular the base voltage, to reverse, thus cutting off the transistor. At this time the inductance of the transformer contains stored energy equal to $\frac{1}{2}L_p i_{pk}^2$, where i_{pk} is the current flowing immediately prior to switching off.

(2.2.2) **Output Stroke.**

The reverse voltage rises until the secondary voltage reaches V_0, the voltage established by previous cycles of operation across the smoothing capacitor C_0. The rectifier D_1 then opens and arrests the secondary voltage at $V_0 + V_{df}$, where V_{df} is the rectifier voltage drop, which is generally a non-linear function of the secondary current i_s. If this is neglected, the secondary current will be seen to decrease linearly according to the equation

$$-L_s\frac{di}{dt} = V_0 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

Its initial value is $i_{pk}N_p/N_s$ (if the effect of stray capacitances and losses occurring during the transient are neglected), so that

$$i_s = \frac{N_p}{N_s}i_{pk} - \frac{V_0}{L_s}t \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

When this current has decayed to zero, the positive voltage across the base winding disappears, and the transistor switches on again with $i_b = I_{bb}$, and i_c rising linearly from zero. The cycle thus recommences.

During the input stroke the transistor operating point (Fig. 2) moves from O steadily along the characteristic $i_b = -I_{bb}$ until it reaches the knee at P. During switch-off it moves rapidly to Q, the point $[v_c = V_i + (N_p/N_s)V_0,\ i_c = I_{c0}]$ on the cut-off characteristic $i_b = +I_{c0}$. It remains at Q for the whole of the output stroke, and then rapidly returns to the origin O.

The duration of the input and output strokes is readily obtained from eqns. (3) and (7) as follows:

$$t_f = L_p\frac{i_{pk}}{V_i} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

$$t_b = \frac{N_s}{N_p}L_p\frac{i_{pk}}{V_0} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

The ratio t_f/T, which will be required later, is thus given by

$$\frac{t_f}{T} = \frac{t_f}{t_f + t_b} = \frac{\frac{N_p}{N_s}V_0}{\frac{N_p}{N_s}V_0 + V_i} = \frac{V_q - V_i}{V_q} \quad . \quad . \quad (10)$$

where $$V_q = V_i + \frac{N_p}{N_s}V_0 \quad . \quad . \quad . \quad . \quad . \quad (11)$$

is the collector voltage of the transistor during the output stroke.

(2.2.3) **Transient Conditions.**

The duration of the switching transients is usually short compared with the times of input and output strokes. It is primarily a function of the ringing frequency of the winding inductances and stray capacitances, but the hole-storage phenomena in the transistor modify the switch-off.

Switching-Off Transient.—The collector current of the transistor does not cease immediately the base voltage goes positive, but decays gradually over a period of a few microseconds, owing to the hole-storage effect in the transistor.[3,6]

This slow current decay somewhat reduces the rate of rise of the voltage across the transformer windings, but the most important consequence is that it is accompanied by a high instantaneous collector dissipation. This is due to the fact that the collector current is still a large fraction of its peak value when the collector voltage has risen to its output-stroke value, V_q. The duration of the decay and the magnitude of the energy loss during the transient tend to vary inversely as the frequency response of the transistor used, but they can be reduced by special circuit measures.*

Switching-On Transient.—At the end of the output stroke the stray capacitances discharge through the transformer windings. After a quarter-cycle of ringing, the transistor base is carried negatively, collector current starts to flow and regenerative switch-on occurs. This transient condition is therefore short, but as will be seen from the following Section, this negative swing of base voltage is normally essential to the maintenance of the relaxation oscillations.

(2.3) Initiation of Oscillations

In many practical circuits, oscillations will not start when the input voltage is applied gradually, unless some negative bias is applied to the base of the transistor. The bias serves to reduce the input resistance of the transistor, which is relatively high at zero base voltage, and thus raises the gain of the feedback loop sufficiently to give instability and regenerative switch-on. Once oscillations are established and an appreciable output voltage has been built up, they will persist owing to the switching-on transient although the bias is removed.

* H. H. van Abbe has recently demonstrated that the transient loss due to the hole-storage effect can be reduced by application of a positive pulse (taken, say, from a tap on the secondary winding via a capacitor) to the base of the transistor during the switching-off transient. In any given convertor, this improvement results in a greater conversion efficiency; it also allows a given transistor type to handle greater power.

The sudden application of input voltage is often sufficient to start oscillations even when no base bias is used. The mechanism is similar to that of the maintenance of oscillations. Shock excitation of the winding self-resonant circuits swings the base of the transistor sufficiently negative for regenerative switch-on to take place.

Whether or not base bias, as shown in Fig. 11, is required to ensure starting in a particular circuit thus depends on the available loop gain and on the switching conditions.

(2.4) Output Voltage and Power Considerations

Before the design of convertor circuits can be discussed, the factors which determine the output voltage must be presented explicitly. It has already been stated that the energy stored in the inductance in each cycle is $\frac{1}{2}L_p i_{pk}^2$; this energy is periodically delivered to the output capacitor, which is assumed to be so large that the voltage across it stays substantially constant over each cycle. The output voltage adjusts itself to an equilibrium value such that the charge periodically fed into the capacitor is equal to the current consumption of the load during the cycle. Disregarding losses, the power fed into the output circuit is $\frac{1}{2}fL_p i_{pk}^2$, where f is the operating frequency. The output voltage is thus given by the equation

$$\frac{V_0^2}{R_L} = \frac{1}{2}fL_p i_{pk}^2 \quad . \quad . \quad . \quad . \quad . \quad (12)$$

This expression, however, is not suitable for design purposes, since f, L_p and i_{pk} are highly interdependent. A more suitable expression for the power, P_0, fed into the output circuit is $\eta V_i I_i$, i.e. the input power times the conversion efficiency, η, which is introduced to allow for losses:

$$\frac{V_0^2}{R_L} = P_0 = \eta V_i I_i = \eta V_i \frac{i_{pk}}{2}\frac{t_f}{T}. \quad . \quad . \quad . \quad (13)$$

Although η must be estimated and t_f/T is a function of circuit parameters [see eqn. (10)], this expression is superior to eqn. (12). It shows clearly that the output power is a linear function of i_{pk} (which can be adjusted by means of R_b) and of the flow-time ratio of the input current, and it is independent of operating frequency and inductance.

The last two parameters, however, have a considerable influence on the losses, and their choice is fully discussed in the following Section.

(3) DESIGN OF D.C. CONVERTOR CIRCUITS

The heart of the d.c. convertor is the transformer, and most of this Section is concerned with its design. The factors governing the choice of the other major circuit components are also considered. The design procedure for optimum efficiency will be treated fully; but it should be noted that by virtue of the tolerance inherent in the circuit acceptable results can be obtained with circuit values quite far from the optima, provided only that the transistor or rectifier is not overstressed thereby.

(3.1) Transformer Design Procedure

The following parameters will normally be prescribed:

(*a*) Input voltage V_i.*
(*b*) Output voltage V_0.
(*c*) Output current I_0.
(*d*) Approximate size of transformer.

* However, if V_i can be freely chosen, it should be made approximately equal to half the maximum collector voltage rating of the transistor to obtain maximum circuit efficiency. This optimum value for V_i is derived in Appendix 13.2.

The following parameters will normally be at the designer's choice:

(*a*) Operating frequency.
(*b*) Core material.
(*c*) Air-gap.
(*d*) Number of primary turns.
(*e*) Turns ratios.

The first four factors are closely interrelated. The operating frequency is the chief parameter; its optimum value is determined by many considerations, one of the most important of which is the core material. The choice of frequency will first be discussed in general terms, and the primary inductance required to realize a certain frequency will be derived. This will be followed by a discussion of the merits of different core materials. The choice of primary/base and primary/secondary turns ratios is independent of the primary circuit design and will be treated last.

(3.1.1) Choice of Frequency.

The correct choice of operating frequency is important if the efficiency is to be high and the transistor dissipation low. The energy losses in the convertor can be divided into those which increase with frequency, those which decrease with frequency and those independent of it.

The main loss which increases with frequency is the transient loss in the transistor. This occurs during switch-off owing to the large collector voltage which exists while considerable hole-storage current flows. Core losses also increase with frequency. They are not of great importance when a ferrite is used, but they must be taken into consideration in iron cores.

The loss which decreases with frequency is the copper loss in the transformer. The operating frequency is nearly inversely proportional to the primary inductance [from eqns. (8) and (9)], so that for a given transformer volume, fewer turns of thicker wire can be used for high-frequency operation. In fact, the copper loss is inversely proportional to the square of the frequency when a particular core is used and the air-gap is adjusted to keep the peak flux density constant.

When it is desired to design for maximum overall efficiency, the optimum frequency, f_0, is given by the minimum in the curve of total loss against frequency shown in Fig. 4. If, on the other hand, it is desired to obtain the maximum output for a given transistor dissipation, ideally a very low frequency should be used. As this will result in large copper losses or a huge transformer, a compromise has to be reached and some frequency below f_0 chosen.

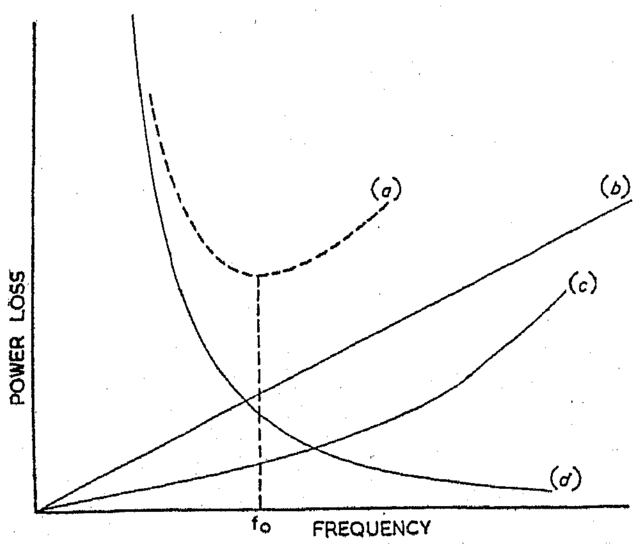


Fig. 4.—Variation of power losses with frequency.
(*a*) Total loss.
(*b*) Transient loss.
(*c*) Core losses.
(*d*) Copper loss.

The actual value of the optimum frequency depends greatly on the transformer size and core material used and also on the magnitude of the transient losses. It can readily be found from preliminary experiments and normally falls in the region of 300 c/s–10 kc/s.

The choice of frequency may also be influenced by factors not connected with efficiency, such as the requirement that the smoothing components should be as small as possible. In this case an operating frequency above f_0 may be adopted.

(3.1.2) **Primary Inductance.**

Once the frequency is fixed, the primary inductance required may be estimated.

From eqn. (8) $$L_p = \frac{V_i}{i_{pk}} t_f$$

$$= \frac{V_i}{i_{pk}} \frac{1}{f} \frac{t_f}{T}$$

From eqn. (13) $$i_{pk} = \frac{2P_0}{\eta V_i} \frac{T}{t_f}$$

Hence $$L_p = \frac{\eta V_i^2}{2P_0} \frac{1}{f} \frac{t_f^2}{T^2}$$

and from eqn. (10) $$= \frac{1}{f} \frac{\eta V_i^2}{2P_0} \frac{(V_q - V_i)^2}{V_q^2} \quad . \quad . \quad . \quad (14)$$

Estimated values of η and V_q must now be inserted. As the normal range of efficiencies is 60–85%, a suitable preliminary estimate for η is often 75%. V_q is a function of V_0, N_s and N_p, but as will be shown in Section 3.1.5, it is normally chosen to have a value V_Q which is somewhat below the maximum peak collector-voltage rating, $V_{c\,max}$, of the transistor used.

The design of the transformer to have the estimated inductance follows established practice.

As the flux is unidirectional, a gap is normally used in all but very-low-power convertors in order to reduce the remanent flux density, B_r,* and to limit the peak flux density to a value, B_{pk}, well out of the saturation region for the core material. When the gap is adjusted so that the latter requirement is met, eqn. (15) (in which B_r is neglected) gives the relationship between L_p, N_p, i_{pk}, B_{pk} and A as follows:

$$L_p = N_p A \frac{B_{pk}}{i_{pk}} \times 10^{-8} \quad . \quad . \quad . \quad . \quad (15)$$

(3.1.3) **Choice of Core Material.**

It follows from eqn. (15) that, provided the core permeability is high enough to allow an appreciable gap to be used, the only core property of major importance is the maximum working flux density of the material. The inductance obtainable for a given geometry and winding is directly proportional to this. Lower optimum frequencies and higher efficiencies may therefore be obtained with materials having a high saturation flux density, since these allow operation at lower frequencies for a given copper loss, which results in lower transient (and other frequency-increasing) losses.

As the optimum operating frequency of d.c. convertors usually falls in the kilocycle-per-second range, ferrite cores are in general the most suitable because of their low eddy-current losses; at high power levels, however, the low frequencies obtainable with iron cores may be advantageous in that the transient losses are substantially reduced. The expected gain in efficiency may, however, not be fully realized owing to increased core losses, particularly if thick laminations are used.

* Alternatively push-pull circuits can be used to give greater transformer utilization. These are also indicated when it is desired to obtain an output power higher than that which can be handled by a single transistor.

(3.1.4) **Number of Base Turns.**

The turns ratio N_b/N_p is determined by the requirement that sufficient voltage must be generated in the base winding to supply a base current, I_{bb}, which is $1/\alpha'$ of the peak collector current required. The required base voltage, V_{bb}, is obtained from curves of V_{bb} versus i_c for the transistor type used. Such curves are shown in Fig. 5 for a transistor having a dissipation rating of 50 mW.

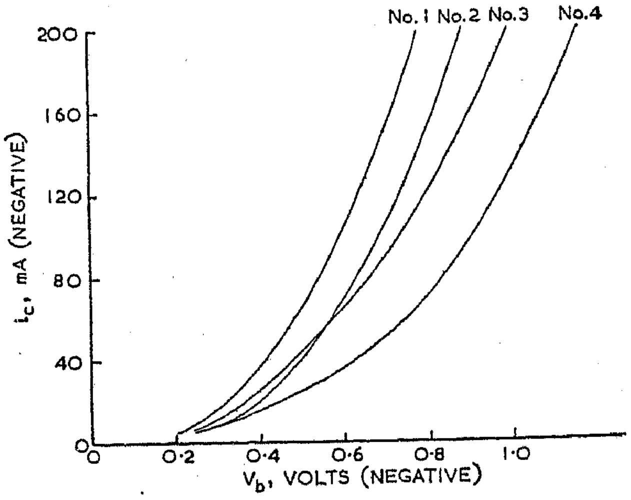


Fig. 5.—Collector current as a function of base voltage.

Curves are plotted for four samples of a transistor type with a dissipation rating of 50 mW.

At the end of the input stroke, the effective voltage across the primary winding of the transformer is given by

$$V_p = V_i - i_{pk}(R_p + R_i + r_c') \quad . \quad . \quad . \quad (16)$$

where R_i is the internal resistance of the battery.

The required turns ratio is thus

$$\frac{N_b}{N_p} = \frac{V_{bb}}{V_i - i_{pk}(R_p + R_i + r_c')} \quad . \quad . \quad . \quad . \quad (17)$$

(3.1.5) **Number of Secondary Turns.**

In principle, the output voltage, V_0, is independent of the primary/secondary turns ratio; in practice, however, the choice of the turns ratio is determined by the requirement that the rated collector-base voltage of the transistor, $V_{c\,max}$, shall not be exceeded during the output stroke.

The total (negative) collector-to-base voltage is

$$V_r = V_i + \frac{N_p}{N_s} V_{spk}\left(1 + \frac{N_b}{N_p}\right) \quad . \quad . \quad . \quad (18)$$

and for $V_r = V_{c\,max}$,

$$\frac{N_s}{N_p} = \frac{(V_0 + V_{df})\left(1 + \frac{N_b}{N_p}\right)^*}{V_{c\,max} - V_i} \quad . \quad . \quad . \quad . \quad (19)$$

It is permissible to use a higher turns ratio, but this is inadvisable from the point of view of circuit efficiency and rectifier inverse rating:

* For this value of N_s/N_p, $V_r = V_{c\,max}$ and $V_q = V_Q$, the maximum permissible collector-emitter voltage in the circuit. Because the analytical expression for V_q is simpler than that for V_r, the limit $V_q = V_Q$ is often used in preference to $V_r = V_{c\,max}$. V_Q can be estimated from

$$V_Q = \frac{V_{c\,max}}{1 + \frac{N_b}{N_p}} \simeq \frac{V_{c\,max}}{1 + \frac{V_{bb}}{V_i}}$$

where V_{bb}/V_i is usually small compared with unity.

(*a*) From eqn. (10), the primary-current flow-time ratio, t_f/T, decreases as N_s/N_p increases. The peak input current must therefore be increased in order to maintain the power input per cycle, with a resultant increase in losses.

(*b*) The rectifier inverse voltage during the output stroke is $V_0 + (N_s/N_p)V_i$, which increases with N_s/N_p.

(3.1.6) Optimum Core Cross-Section.

There is an optimum copper/core ratio which is of the same order as that of ordinary transformers working at the same frequency. It is not critical, so that efficient operation can be obtained for a considerable range of ratios. It should be remembered that, for constant inductance and constant flux density in the core (i.e. varying air-gap), the number of turns required varies inversely as the core cross-section [from eqn. (15)].

(3.1.7) Primary/Secondary Winding-Space Ratio.

There is an optimum primary/secondary winding-space ratio for which the total copper loss is a minimum. With the simplifying assumption that the average turn lengths of the primary and secondary windings are the same, and with no allowance for insulation space, the fraction k of the winding space that should be allocated to the primary winding is given by

$$k = \frac{1}{\sqrt{\left(\frac{N_s}{N_p}\frac{V_i}{V_0}\right)} + 1} \quad . \quad . \quad . \quad . \quad . \quad (20)$$

This expression is derived in Appendix 13.1.

(3.1.8) Limitations of the Foregoing Design Procedure.

The design recommendations given form the basis of the design of all convertors having a simple rectifier output.

For very-low-power convertors, however, great weight must be given to factors which have not been discussed, because their general importance is small at higher powers. Thus losses which are independent of power level should be kept low. These include losses due to reverse currents in rectifiers and transistor junctions, and the energy lost in charging and discharging stray capacitances.

In high-voltage units difficulties can arise because of the very high turns ratios required. The use of voltage-multiplier systems, the design of which is dealt with below, is then indicated.

(3.2) Choice of Components

(3.2.1) The Transistor.

The circuit can be designed to work with junction transistors of almost any characteristics, although in the interests of efficiency it is preferable for the transistor to have a low "on" resistance, r'_c, a high collector rating, V_{cmax}, and a fast hole-storage decay. It is essential to choose a type with an adequate dissipation rating. This must be based on an analysis of the conversion losses, for the fraction of the total power handled which is dissipated in the transistor may vary widely depending on circuit conditions. This fraction may well be as low as one-tenth.

(3.2.2) Rectifier.

The types of rectifier which are normally used in d.c. convertors are germanium junction diodes, point-contact diodes or selenium rectifier stacks. The most efficient of these is the junction rectifier with its low forward drop and small reverse current, but it is at present available in a few ratings only. Point-contact diodes are satisfactory at output voltages of a few tens of volts and currents of the order of 1–50 mA. Higher output voltages than those possible with a single unit of either type cannot be obtained by connecting two or more rectifiers in series, unless they are selected to have their reverse characteristics sufficiently similar so that the reverse voltage is shared approximately equally. Selenium rectifiers are also suitable and are preferable for very low output currents ($\leqslant$1 mA) and for reverse voltages higher than those handled by the germanium types.

The average forward current in the rectifier is I_0, and its peak value is given by

$$2I_0\left(1 + \frac{N_p}{N_s}\frac{V_0}{V_i}\right)$$

The reverse voltage across the rectifier is given by $V_0 + (N_s/N_p)V_i$, and it exists during the whole of the input stroke. The semiconductor rectifier chosen must be capable of withstanding the reverse voltage when applied for a substantial portion of the time.

(3.2.3) Series Base Resistance.

The resistor, R_b, controls the power input to the circuit and hence the output. For any given transistor the output power may be adjusted by means of R_b; alternatively, if a fixed output power is required, R_b may be used to set up the circuit using different transistors. In the latter case a variable resistor must be provided having a value equal to

$$\frac{\alpha'_{max}(V_{bw} - V_{bbmin})}{i_{pk}}$$

where α'_{max} and V_{bbmin} are, respectively, the highest and lowest values pertaining to any transistor. As α' varies somewhat with i_c, its value for $i_c = i_{pk}$ must be used.

The value of R_b used should not be greater than that required for any particular set of circumstances, since the duration of hole-storage currents and losses increases with R_b.

(4) PERFORMANCE OF TYPICAL CIRCUIT

A high-power convertor working from a 12-volt input and giving 4 watts output will be taken as an example. A development-type power transistor capable of dissipating some 2–3 watts was used in this circuit with germanium junction power rectifiers in the output. Two transformers were used, one having a ferrite core and operating at 1 200 c/s, the other being a silicon-iron C-type (0·004 in strip) core and operating at 700 c/s.

(4.1) Efficiency

Efficiencies of 70% and 75% respectively were obtained from circuits using the above types of core.* The main sources of power loss are the copper losses in the transformer (which amount to some 25% of the total loss), the loss in the collector resistance r'_c (which accounts for about 20% of the losses), and the transient loss in the transistor. The latter is very frequency-dependent and may amount to 30% of the total loss when a ferrite core is used. The remaining losses, i.e. those in the base circuit, in the rectifier, the transistor back losses, and the core losses in the transformer are individually relatively small, although collectively they account for the remaining 20–30% of the total power loss.

The relative importance of the losses will, of course, not be the same in all convertors.

For example

(*a*) At low powers, the losses independent of power level, mentioned in Section 3.1.8, become relatively more important.

(*b*) In most transistors the decay of hole-storage current is more rapid than in the experimental power transistors used in the 4-watt circuit, so that transient losses will normally be relatively lower, and the optimum frequency higher.

(*c*) The rectifier losses will normally be appreciable when germanium junction rectifiers cannot be used.

* All data in this paper apply to circuits in which no measures have been taken to reduce transient losses.

(4.2) Loss in Battery

An additional power loss of practical importance is the loss in the internal impedance, R_i, of the battery, given by

$$\tfrac{1}{3}i_{pk}^2 R_i \frac{t_f}{T}$$

If the battery voltage is kept constant over the cycle by means of a large smoothing condenser across it, this may be reduced to

$$\tfrac{1}{4}i_{pk}^2 R_i \left(\frac{t_f}{T}\right)^2$$

(4.3) Effect of Ambient Temperature

The reverse currents both in the transistor and the rectifier increase with temperature, so that a fall in efficiency may be expected as the ambient temperature rises. A reduction in the forward resistance of both semi-conductor components occurs, however, which partially cancels the effect of the increased reverse losses.

Measurements on a 4-watt convertor between 20° and 40° C have shown that a fall in output voltage of 1–2% occurs owing to a fall in efficiency of 2–3% over this temperature range.

(4.4) Effect of Input-Voltage Changes

The output voltage is proportional to the input voltage when the load is resistive. When the load is non-linear the output power is approximately proportional to the square of the input voltage.

(4.5) Output Impedance

If the power output of the circuit were absolutely constant irrespective of the load, R_L, the output impedance of the circuit, R_0, would always be exactly equal to R_L (this follows from the converse of the optimum power transfer theorem, or it can be derived algebraically). In fact, the output power of the d.c. convertor varies only slightly with the load; a fall in output power P_0 occurs at low output voltages because of a reduction in input power caused by mark/space ratio changes. Because of this power variation, and of minor changes in losses with output voltage, R_0 is only approximately equal to R_L.

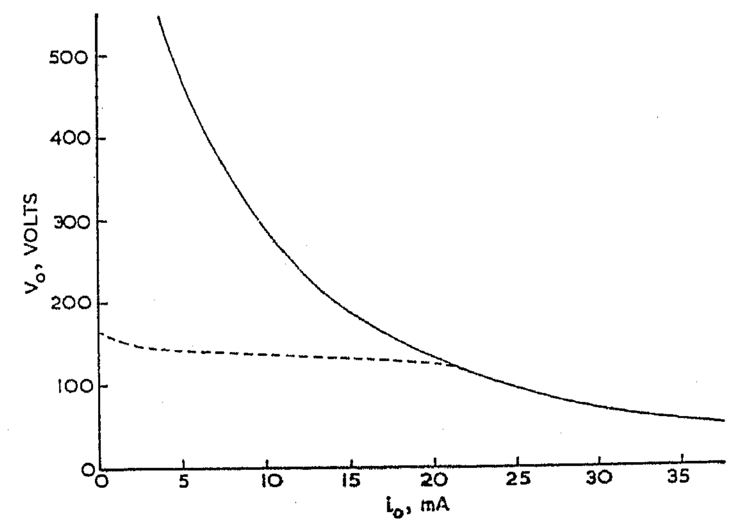


Fig. 6.—Regulation of a d.c. convertor operating at a power level of 2·5 watts approximately.

——— Regulation of basic circuit.
– – – – Regulation obtained with circuit incorporating a simple stabilization system such as that shown in Fig. 8 or Fig. 9.

Fig. 6 shows the regulation of a high-power convertor under conditions of variable load. (Special precautions were taken to avoid component or transistor breakdown at the extremes of the characteristic.)

(5) STABILIZATION OF OUTPUT VOLTAGE

When the load or the input voltage may vary and the consequent changes in V_0 are unacceptable, some stabilizing circuit must be added. This may be a conventional one external to the convertor, or it may take the form of extra circuits added to the convertor itself which improve its output characteristics. The former (gas-filled stabilizers or voltage-dependent resistors) are intrinsically wasteful of power. Internal stabilization, however, can give more economical operation by returning excess power to the battery or adjusting the power drawn from it to the load requirements. Some feedback systems which accomplish the latter will be described.

(5.1) Feedback Stabilizing Systems

(5.1.1) General Principles.

Feedback systems which apply control to the base circuit of the switching transistor can take their input either from the output line or from points in the circuit which periodically reach a voltage which is a measure of the output voltage. As the loading imposed by the feedback circuit on the output line may be excessive, the second alternative is often preferable. The voltage existing during the output stroke on a tap on the secondary winding is proportional to the output voltage, if the voltage drops in the rectifier and secondary resistance (which are normally small) are neglected. If a moderate degree of stabilization is acceptable, feedback may be taken from such a tap or from a separate winding. It can be applied to the switching transistor as current or voltage bias. In either case it is phased so as to reduce the base current, I_{bb}, drawn during the forward stroke, if the output voltage rises above the nominal value. The transistor then switches off at a lower peak collector current and reduces the input power drawn.

In practice, steps must be taken to ensure that the normal relaxation oscillations of the convertor are not interfered with by the stabilizing circuit, and that feedback is effective during the input stroke, even though it may be derived from a voltage existing only during the output stroke.

The three feedback stabilizing circuits to be described will serve as examples of a variety of possible arrangements.

(5.1.2) Specific Circuits.

In Fig. 7 a feedback chain comprising transistors T_3 (as an impedance-matching stage) and T_2 (as a voltage comparator and

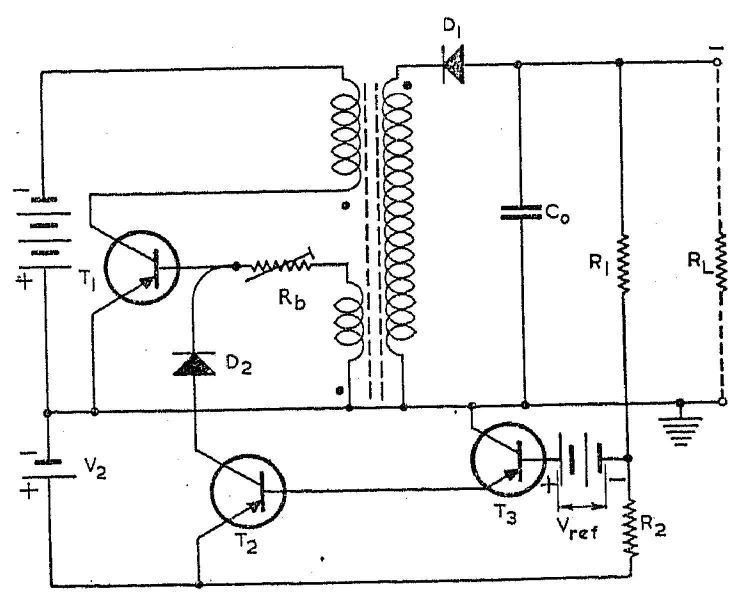


Fig. 7.—Convertor with feedback circuit to stabilize the output voltage.

current amplifier) is connected between a voltage divider R_1, R_2 across the output and the base of the switching transistor T_1, which is fed from a higher value of R_b than usual. As the negative output voltage rises above its nominal value, which is

$V_{ref}(R_1 + R_2)/R_2$ approximately, T_2 draws an increasing current through R_b and thus decreases the base current, I_{bb}, of T_1 during the input stroke.

This type of circuit is only suitable for convertors with moderately low voltage step-up ratios.

The stabilizing circuit of Fig. 8, although intrinsically not so

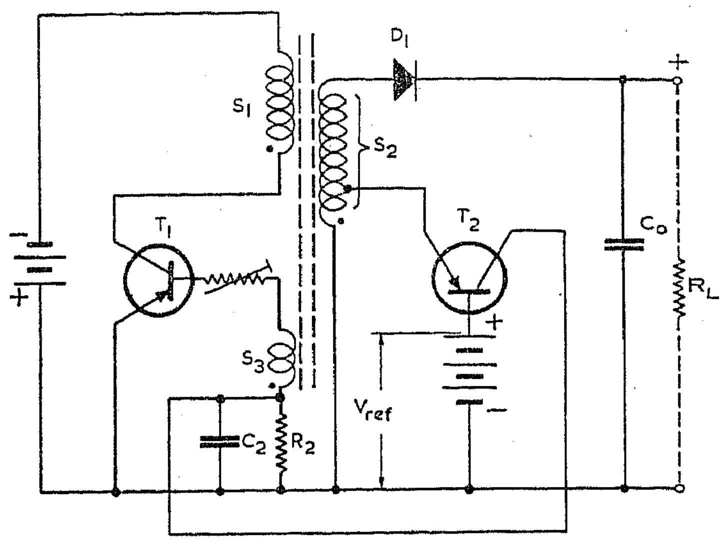


Fig. 8.—Convertor fitted with another type of stabilizing circuit.

effective, can be applied more generally. The tap on the secondary winding, S_2, is positioned so that when the wanted output voltage appears across S_2, the tap voltage slightly exceeds the reference voltage, V_{ref}. The bulk of the emitter current which therefore flows in the gate transistor, T_2, during the output stroke goes to the collector, and builds up a positive bias voltage across the load R_2C_2. If the output voltage rises for any reason, the bias increases considerably. This reduces the base current drawn by T_1 during the input stroke, and thus reduces the rise in output voltage.

The reference battery is charged by a small current $[1/(\alpha' + 1)$ of the emitter current of $T_2]$ and can with advantage be replaced by a germanium reference diode (Zener effect diode).

The circuit of Fig. 9 stabilizes the output voltage against

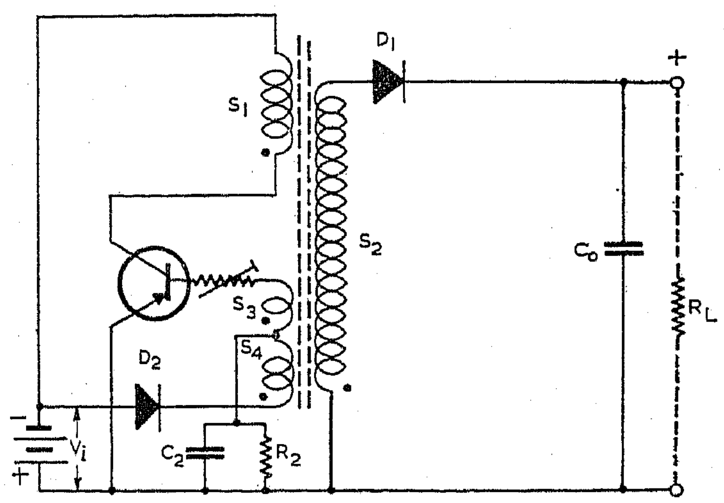


Fig. 9.—Convertor with feedback circuit to improve regulation.

variations of load, but not of input voltage. Its action is similar to that of the last circuit. The diode D_2 which is backed off by the input battery acts as a gate, so that the current in the feedback loop is proportional to the amount by which the voltage on the feedback winding S_4 exceeds V_i during the output stroke.

(6) OVER-VOLTAGE PROTECTION CIRCUIT

In many cases where stabilization of the output voltage is not necessary, it is still desirable to protect the convertor and equipment against damage due to the generation of an excessive

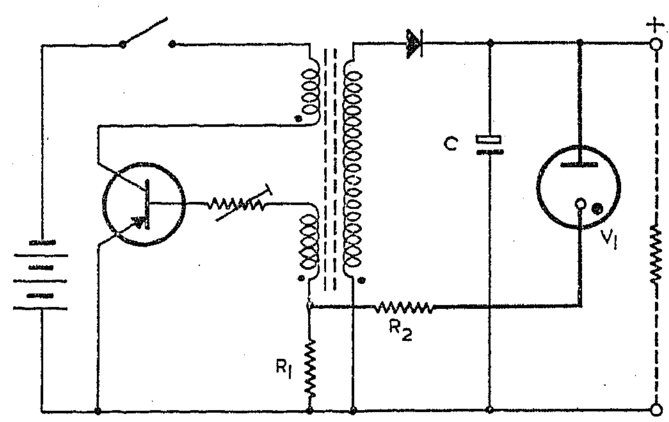


Fig. 10.—Convertor fitted with protection circuit.

If the output rises above the striking voltage of the gas-filled diode V_1, the circuitis rendered quiescent.

voltage if the load becomes open-circuited or too light. Fig. 10 shows a simple protective circuit which switches off the convertor if, for any reason, the positive output voltage reaches some predetermined voltage above the normal output level.

It has already been pointed out that the relaxation-oscillator circuit is stable (or can readily be made stable by increasing the loading on the transformer) in the static condition of zero base voltage and the transistor almost cut-off, but that oscillations can nevertheless often be started simply by switching on the input voltage, and they are then self-maintaining. If, under these conditions, the oscillations are interrupted by a positive pulse on the base, and this is allowed to decay slowly, they will not start again. The gas-filled diode, V_1, and resistor R_1 serve to apply such a pulse. The diode V_1 is chosen to have a striking voltage, V_s, above the normal output voltage. If, for any reason, the output voltage rises to V_s, breakdown takes place and the voltage across the gas-filled diode falls to a lower value, i.e. its burning voltage V_m. The difference $(V_s - V_m)$ appears across R_1 and R_2 as a positive pulse of several volts with a decay time-constant $C_0(R_1 + R_2)$, which can readily be made long enough to prevent switching on again. The circuit can be reset from its shut-off state by switching off the input and switching on again. It will then oscillate and function normally if the fault has been cleared, or shut off again if not.

(7) VOLTAGE-MULTIPLIER CIRCUITS

When high output voltages are required, a very high turns ratio is necessary in the simple circuit. This leads to practical difficulties in providing good coupling between primary and secondary windings, and to large winding capacitances which hold much energy. The use of voltage-multiplying systems in keeping the turns ratio to a minimum can therefore be very advantageous.

(7.1) General Considerations

Two commonly used voltage-doubler arrangements are shown in Fig. 11. When used with the d.c. convertor they are not strictly voltage-doubling systems, since the voltages developed across the secondary winding are not necessarily equal during the two parts of the cycle.

The main difference from the simple convertor circuit is that during the input stroke there is now a throughput of energy into the output circuit, with a consequent change in input-current waveform.

The contribution to the output voltage developed during the input stroke* will be referred to as V_T. It is determined by simple transformer action

$$V_T \simeq V_i \frac{N_s}{N_p} \quad \quad (21)$$

* The terms "input stroke" and "output stroke" will be retained for the description of voltage-doubler circuits, although the relevance of the terms is now confined to that part of the power which is stored.

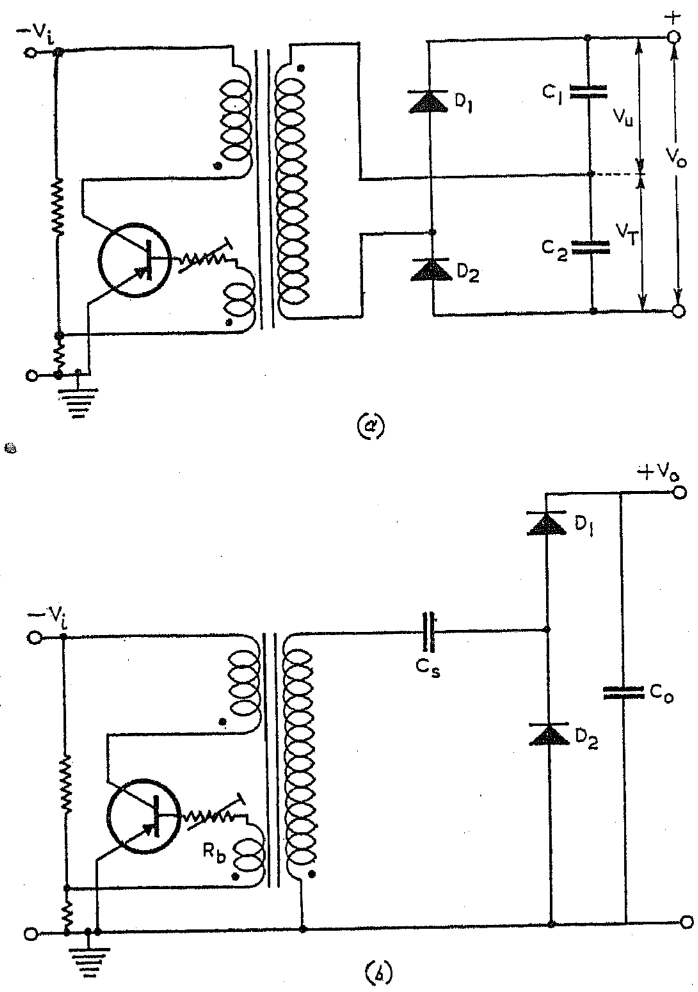


Fig. 11.—Convertors with voltage-doubler output arrangements.

(a) With symmetrical circuit.
(b) With diode-pump circuit.

The balance of the output voltage is derived as before from the magnetic energy stored in the inductance. It is delivered to the output circuit during the output stroke, and its magnitude, which will be referred to as V_U, is a function of the loading.

The total output voltage, $V_0 = V_T + V_U$, is thus also a function of the load, but the regulation of the voltage-doubler circuit is better than that of the simple circuit because of the fixed nature of V_T.

The transformer loading during the input stroke modifies the primary-current waveform as shown in Fig. 12. It follows from

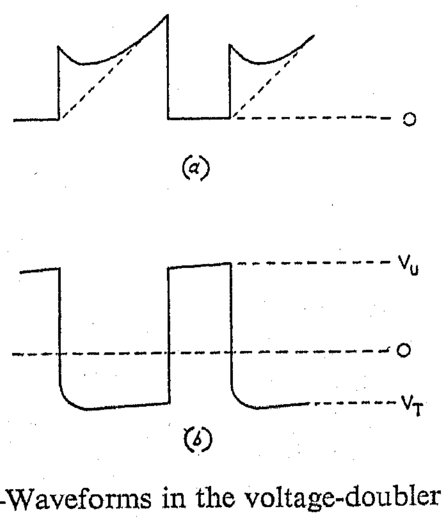


Fig. 12.—Waveforms in the voltage-doubler circuits.

(a) Input current.
(b) Voltage across secondary winding.

energy considerations that the ratio of the area above the dotted line to that below is equal to the ratio V_T/V_U. This ratio should not exceed approximately one-half, for otherwise

(a) The heavy loading on the transformer makes it very difficult to start oscillations even when negative base bias is used.

(b) The transistor cannot carry the current peak at the beginning of the input stroke without coming out of bottoming and thus introducing an extra loss.

With V_T/V_U equal to one-half, the turns ratio N_s/N_p in a voltage-doubler circuit can be reduced to two-thirds of that required in the simple circuit for a given output voltage, and to one-third in a voltage quadrupler.

(7.2) Symmetrical Voltage-Doubler Circuit

In the circuit of a symmetrical voltage-doubler, which is shown in Fig. 11(a), the two condensers C_2 and C_1 are separately charged to $-V_T$ and V_U during the input and output strokes respectively. The output voltage is then given by

$$V_0 = V_U + V_T \quad . \quad . \quad . \quad . \quad . \quad . \quad (22)$$

(7.3) Diode-Pump Voltage-Doubler Circuit

The diode-pump voltage-doubler circuit is of greater interest as it also forms the basis for higher-order multiplication. The circuit is shown in Fig. 11(b). During the input stroke, energy is stored in the condenser C_s, which is charged to the voltage V_T by conduction in D_2. At the beginning of the output stroke, the secondary voltage rises to V_U, and the rectifier D_2 cuts off. Current then flows through C_s and D_1 into the output capacitor C_0, and the output voltage, neglecting the drop in the rectifier, is given by

$$V_0 = V_U + V_T$$

For efficient operation of the circuit, C_s must be so large that the voltage across it remains substantially constant, and equal to V_T, throughout the output stroke. The secondary voltage then remains substantially equal to V_U during this time. The operation is then the same as that of the simple circuit except that V_0 must be replaced by $(V_0 - V_T)$ in equations such as eqns (7) and (19).

It may be shown that to attain this mode of operation C_s must satisfy the following condition:

$$C_s \gg \frac{1}{2}\frac{L_p i_{pk}^2}{V_T(V_0 - V_T)} \quad . \quad . \quad . \quad . \quad . \quad (23)$$

If this condition does not hold, there will be an undesired rise of secondary voltage during the output stroke from $(V_0 - V_T)$ towards, or even beyond, V_0.

The operating principles of voltage tripler and higher-order multiplier circuits are similar to those of the diode pump.

(8) APPLICATIONS OF D.C. CONVERTORS

Convertors using the circuit principles outlined above can be made to a wide range of specifications. With the transistors presently available and shortly forthcoming, it is possible to make convertors with output ratings of several watts and output voltages up to several kilovolts, working from input voltages of a few volts only.

Five examples will serve to illustrate the variety of requirements that convertors are capable of meeting. The types of component used and the efficiency obtained in each unit are listed below.

(a) Miniature milliwatt convertor for supplying two hearing-aid pre-amplifier valves, or one semi-electrometer valve, with h.t. power.

Input	1·3 or 2·6 volts.
Output	30 volts, 70 μA.
Transistor	Low-level amplifier, 25 mW rating.
Rectifier	Miniature selenium.
Transformer ..	Dimensions: 1·3 × 1·1 × 1·0 cm. Ferrite core.
Efficiency	60% with 2·6 volts input.

(*b*) H.T. generator for frequency-changer and intermediate-frequency-amplifier valves in a battery radio set with transistor a.f. stages.

Input	6 volts.
Output	45 volts, 3 mA.
Transistor	50 mW rating.
Rectifier	Germanium point-contact diode.
Transformer	Dimensions: 4 × 4 × 2·5 cm. Ferrite core.
Efficiency	80%.

(*c*) H.T. generator for heavy-current Geiger–Müller counter.

Input	4·5 volts.
Output	500–700 volts, 50–35 μA.
Transistor	50 mW rating.
Rectifier	Selenium in voltage-quadrupler circuit.
Transformer	Dimensions: 3·5 × 2·5 × 3·0 cm. Ferrite core.
Efficiency	70%.

(*d*) Four-five watt h.t. convertor suitable for small portable transmitters and receivers.

Input	12 volts.
Output	100–150 volts.
Transistor	2 watts rating.
Rectifier	Germanium junction diode.
Transformer	Dimensions: 5 × 5 × 3 cm, Silicon-iron C core, 0·004 in laminations.
Efficiency	75%.

(*e*) H.T. generator for demonstration battery-operated oscillograph system.

Input	12 volts.
Output	(i) 2 kV, 0·8 mA for cathode-ray-tube and potentiometer circuits. (ii) 150 volts, 3 mA for simple valve amplifier.
Transistor	2 watts rating.
Rectifier	Selenium. Voltage quadrupler circuit for 2 kV output. Simple rectifier from tap on secondary winding for 150-volt output.
Transformer	Dimensions: 7 × 6 × 5 cm. Ferrite core.
Efficiency	70%.

Efficiencies of up to 90% can often be obtained if the input voltage is favourable, the transformer has generous dimensions, a transistor type of ample ratings is used and an efficient rectifier is available.

It will be seen from Fig. 13, which shows units (*a*) and (*d*), that the size and weight of transistor d.c. convertors can be very moderate. This, coupled with their high efficiency, should make their adoption advantageous in many applications where h.t. batteries are required at present. For example, a great saving in weight, volume and running cost is obtained when the e.h.t. supply for a Geiger–Müller tube or photo-multiplier is obtained from a few small l.t. cells and a transistor convertor, or if a convertor allows the h.t. requirements of a small portable transmitter to be drawn from the accumulator which normally supplies the heater power only.

(9) COMPARISON OF TRANSISTOR D.C. CONVERTORS WITH OTHER VOLTAGE STEP-UP DEVICES

Of the devices listed in the Introduction, only the vibrator and the rotary convertor are suitable for use with l.t. supplies. Both become increasingly inefficient as the output required is reduced, so that at powers of less than about 1 watt there is no important rival to the transistor d.c. convertor. At higher power levels, however, the vibrator power supply competes with the transistor convertor. Although more operating experience on the transistor oscillator is required before the competitive situation can be assessed, the following seem likely to be the main points of difference.

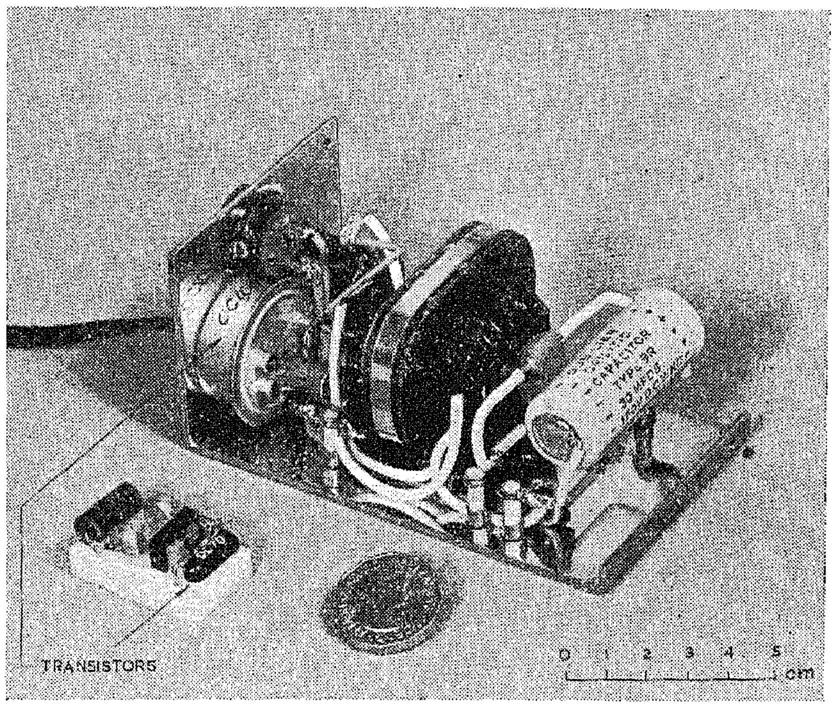


Fig. 13.—2 mW and 4-watt convertor units.

The transistor convertor is expected to be superior in size, weight, reliability, life and freedom from arc-generated interference, whereas the vibrator supply is superior mainly in the absence of temperature limitations, regulation, and (at least at present) in initial cost and power-handling capacity. The working efficiency of the transistor circuit is higher below a certain power level, which is likely to be in the region of several watts.

(10) CONCLUSIONS

Transistor d.c. convertors, such as the circuits described in detail in the paper, offer a convenient and efficient solution to the problem of generating high voltages from batteries at power levels up to several watts. They fill a long-existing need for practical ways of doing this at power levels in the milliwatt and fractional-watt range, and promise to be superior to existing methods at powers of a few watts on the counts of size, weight, reliability, life and efficiency. They are, however, subject to the usual temperature limitations associated with semi-conductor devices, but these will not be serious when silicon devices become generally available.

(11) ACKNOWLEDGMENTS

The authors wish to thank their colleagues, Mr. G. O. Crowther, for contributing the calculation of optimum input voltage; Messrs. H. H. van Abbe, A. Dorn, C. F. Hill, P. A. Neeteson and E. Wolfendale, for valuable discussions; and Mr. A. J. Heins v. d. Ven, for helpful advice on the preparation of the paper. They are indebted to the Directors of the Mullard Radio Valve Co., Ltd., for permission to publish the paper.

(12) REFERENCES

(1) Bryan, G. W., Jr.: "Application of Transistors to High Voltage Low Current Supplies," *Proceedings of the Institute of Radio Engineers*, 1952, **40,** p. 1521.

(2) Cooke-Yarborough, E. H., Stephen, J. H., and James, J. B.: British Patent Application No. 20668, 1953.

(3) Ebers, J. J., and Moll, J. L.: "Large Signal Behaviour of Junction Transistors," *Proceedings of the Institute of Radio Engineers*, 1954, **42,** p. 1761.

(4) PEARLMAN, A. R.: "Transistor Power Supply for Geiger Counters," *Electronics*, 1954, **27**, p. 144.
(5) JOHNSTON, D. L.: "Transistor H.T. Generator," *Wireless World*, 1954, **60**, p. 518.
(6) MOLL, J. L.: "Large Signal Transient Response of Junction Transistors," *Proceedings of the Institute of Radio Engineers*, 1954, **42**, p. 1773.

(13) APPENDICES

(13.1) Derivation of Optimum Primary/Secondary Winding-Space Ratio

Resistive loss in primary winding is

$$\frac{1}{3}i_{pk}^2\frac{t_f}{T}R_p$$

Resistive loss in secondary winding is

$$\frac{1}{3}i_{pk}^2\frac{N_p^2}{N_s^2}\frac{t_b}{T}R_s$$

Therefore the total copper loss, P_t, is

$$\frac{1}{3}i_{pk}^2\frac{t_f}{T}\left(R_p + \frac{N_p}{N_s}\frac{V_i}{V_0}R_s\right)$$

Let k = fraction of winding space occupied by primary winding,
and R = resistance that a winding of one turn occupying the whole window would have.

Then, if any variation in average turn length and the space taken by insulation between turns is neglected, the resistance of a winding of N turns occupying $1/k$ of the window is RN^2/k.

Therefore $$P_t = \frac{1}{3}i_{pk}^2\frac{t_f}{T}R\left[\frac{N_p^2}{k} + \frac{N_sN_p}{(1-k)}\frac{V_i}{V_0}\right] \quad . \quad . \quad . \quad (24)$$

The minimum value of this expression occurs when

$$\frac{dP_t}{dk} = 0$$

i.e. when $$k = \frac{1}{\sqrt{\left(\frac{N_sV_i}{N_pV_0}\right)} + 1} \quad . \quad . \quad . \quad . \quad . \quad . \quad (25)$$

(13.2) Derivation of Optimum Input Voltage

The effect of the variation of V_i on various losses will be examined under practical conditions.

The primary requirement is that an output of V_0I_0 is to be achieved with a transformer of given core and window area. (A gapped core is assumed.)

Hence $$P_i = \frac{V_0I_0}{\eta} = V_iI_i = V_i\frac{i_{pk}}{2}\frac{t_f}{T} \quad . \quad . \quad . \quad . \quad . \quad (26)$$

(As any variation of η with V_i, which has only second-order effects, will be neglected, P_i is assumed to be constant.)

The following secondary requirements must be laid down to give a fair basis for examination of the effects of varying V_i.

V_q, B_{pk} and T must be kept constant and equal to V_Q, B_1 and T_1 respectively, where V_Q, as in Section 3.1.5, is somewhat below the maximum collector-voltage rating of the transistor, B_1 is the maximum advisable flux density in the core, and T is kept constant so that frequency variations do not exert an obscuring influence.

Hence from eqn. (11),

$$V_i + \frac{N_p}{N_s}V_0 = V_Q \quad . \quad . \quad . \quad . \quad . \quad . \quad (27)$$

From eqn. (15),

$$\frac{L_pi_{pk}}{10^{-8}AN_p} = \frac{L_pi_{pk}}{CN_p} = B_1 \quad . \quad . \quad . \quad . \quad . \quad . \quad (28)$$

And from eqns. (8) and (9),

$$L_pi_{pk}\left(\frac{1}{V_i} + \frac{N_s}{N_p}\frac{1}{V_0}\right) = T_1 \quad . \quad . \quad . \quad . \quad (29)$$

where $C = 10^{-8}A$ is a constant.

It may be shown that eqns. (26)–(29) will be satisfied, if the remaining five variables i_{pk}, t_f, N_p, N_s and l_a are the following explicit functions of the independent variable V_i:

$$i_{pk} = \frac{2P_iV_Q}{V_i(V_Q - V_i)} \quad . \quad . \quad . \quad . \quad . \quad . \quad (30)$$

$$N_p = \frac{T_1}{CB_1}\frac{(V_Q - V_i)V_i}{V_Q} \quad . \quad . \quad . \quad . \quad . \quad (31)$$

$$N_s = \frac{T_1}{CB_1}\frac{V_iV_0}{V_Q} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (32)$$

$$t_f = T_1\left(1 - \frac{V_i}{V_Q}\right) \quad . \quad . \quad . \quad . \quad . \quad . \quad (33)$$

$$l_a = \frac{8\pi P_iT_1}{10CB_1^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (34)$$

The various losses can now be expressed as functions of V_i.

Consideration of Losses.

(*a*) *Losses, P_r, in the forward resistance of the bottomed transistor r'_c.*

$$P_r = \tfrac{1}{3}r'_ci_{pk}^2\frac{t_f}{T}$$

$$= \frac{4}{3}r'_c\frac{P_i^2}{V_i^2}\frac{V_Q}{(V_Q - V_i)} \quad . \quad . \quad . \quad . \quad . \quad . \quad (35)$$

This has a minimum when $V_i = \tfrac{2}{3}V_Q$.

(*b*) *Transient losses, P_h, due to hole storage.*

As T and V_q are kept constant, the only important factor influencing this loss is i_{pk}. P_h is a minimum when i_{pk} is a minimum. Therefore, from eqn. (30) P_h has a minimum when

$$V_i = \tfrac{1}{2}V_Q \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (36)$$

(*c*) *Loss in the transformer.*

If the values for N_p and N_s from eqns. (31) and (32) are inserted in eqn. (24) the following expression for the total copper loss, P_t, is obtained:

$$P_t = \frac{4}{3}\frac{RT_1^2P_i^2}{C^2B_1^2}\left[\frac{1}{k}\left(1 - \frac{V_i}{V_Q}\right) + \frac{1}{(1-k)}\frac{V_i}{V_Q}\right] \quad . \quad (37)$$

If the absolute optimum design is required, k should be treated as a function of V_i [that given by eqn. (25)]. For the sake of simplicity, however, k will be treated as a constant and equal to one-half. It will be shown later that this value of k is indeed near the optimum.

For $k = \tfrac{1}{2}$, $$P_t = \frac{8}{3}\frac{RT_1^2P_i^2}{C^2B_1^2} \quad . \quad . \quad . \quad . \quad . \quad (38)$$

and is independent of V_i.

(*d*) *Overall minimum of losses.*

For $k = \frac{1}{2}$, therefore, the optimum value of V_i is given by the consideration that the two losses previously considered, P_r and P_h, should be at their minimum,

i.e. $$\tfrac{1}{2}V_Q < V_i < \tfrac{2}{3}V_Q$$

In practice, V_i will not be made greater than $\frac{1}{2}V_Q$, since the rectifier inverse voltage rating, given by $V_0[1 + V_i/(V_Q - V_i)]$, increases rapidly with V_i beyond $V_i = \frac{1}{2}V_Q$, and with it increase the rectifier cost and losses. For $V_i = \frac{1}{2}V_Q$, the value $k = \frac{1}{2}$ previously assumed is indeed the optimum.

Thus $V_i \simeq \frac{1}{2}V_Q$ is normally the optimum value for maximum overall efficiency. As V_Q is only a little less than V_{cmax}, $V_i \simeq \frac{1}{2}V_{cmax}$ is the optimum value of V_i. This optimum is independent of factors such as V_0 and P_0.

DISCUSSION ON THE ABOVE FOUR PAPERS, AND ON THE PAPER BY MR. G. R. NICOLL,* BEFORE A JOINT MEETING OF THE RADIO AND MEASUREMENTS SECTIONS, 11TH MAY, 1955

Dr. J. R. Tillman: Mr. Nicoll's conclusions are definite and convincing, but some of the theory he quotes may need revision. In eqns. (4)–(13), he introduces ϕ, the barrier height, and its dependence on applied voltage, $d\phi/dV$. Are these quantities deduced from the i/V relationships? If so, can the relationships satisfy, simultaneously, both eqn. (4) which, if $\phi = \phi_0 + k_1V + k_2V^2 \ldots$ can be rewritten

$$i = i_0[\exp(-e\phi_0/kT)]\{\exp[-(k_1V + k_2V^2\ldots)e/kT]\}[\exp(eV/kT) - 1]$$

and the equation of Section 4, $i = A(\exp \alpha V - 1)$ where α is found to be, not $e/kT = 39V^{-1}$ at room temperature, but something less, e.g. 25 volt^{-1} for diode (*c*) and 16 volt^{-1} for diode (*d*). For germanium point-contact diodes, and using theories which also yield eqn. (4), measurements of the i/T dependence at very small biases and of the capacitance-voltage dependence enable values of ϕ to be deduced. Agreement between the values obtained is poor. Is it any better for silicon?

Mr. Stephenson draws attention to the role of generator impedance in determining noise factor, and endeavours to present the engineer with data whose use demands no knowledge of the mechanisms responsible for noise. He finds a frequency dependence (see Fig. 8) which differs markedly from that obtained if one generator is placed at the output—where other evidence would put it. The reduction of the noise parameters V_g and I_g with increasing frequency is arrested at about 15 kc/s and is converted into an increase above about 25 kc/s, largely because α_{cb} for the type of transistor used has fallen 3 dB at about 15 kc/s and a further 6 dB at about 30 kc/s.

The photo-electric cell using a diffused junction is hermetically sealed by solder—a technique used in some other crystal valves. Are Messrs. Waddell, Mayer and Kaye satisfied that it is a sound manufacturing prospect? The index they find in the dependence of noise on frequency, namely $-1\cdot5$, is greater than the usual. The points of their Fig. 3 corresponding to all but the three lowest values of noise current measured suggest an index of about $-1\cdot25$. Only if the load was 100 kilohms or more and its contribution to the noise measured was accurately known do the lower points seem worthy of equal weight.

The extension of the long-wave cut-off to 2 microns is very interesting; but is the evidence which excludes the authors' first explanation (that the energy gap to the conduction band from the acceptor level in the gold-doped *p* layer, $0\cdot60$ eV, corresponds well with sensitivity to 2 microns) strong enough? If the spot of "light" used to measure the curve for the *n* side in Fig. 5 was focused only 100 microns (i.e. about one-half of a diffusion length) from the junction and came from a cone of light of half-angle 45°, the cone should be transmitted in the *n* side with a half-angle of about 10°. But the cell is 800 microns deep, so that some of the cone would enter the *p* region directly; internal reflections, pronounced because of the high refractive index, would contribute also to the "light," which, although entering the *n* side, reaches the *p* side.

The paper by Messrs. Hilbourne and Jones gives a useful analysis of the subject. Table 1 would benefit from another column showing the harmonic production, so that the amount of feedback needed (and hence the reduction in gain to be faced) to reduce the distortion to some common figure could be calculated. I am sorry to see two symbols, α_o and β, well established in transistor contexts as representing current gain at low frequencies and the base transmission factor, respectively, given other meanings; the shortage of symbols is acute, but first usages, even when not the best, should be amended only by agreement. The quoted rise of 1° C in junction temperature per 5 mW of collector dissipation for a low-power transistor is a very acceptable figure; can it be much reduced by the fitting of an external heat sink? The dependence of α on I_e is not well explained. At low currents surface recombination is apparently primarily responsible for $(1 - \alpha)$; as the current rises, the increased base current sets up an appreciable drift field which, according to Webster, tends to divert the injected carriers from the surface so that a smaller percentage recombines there. At still higher currents the density of minority carriers, and hence the density of majority carriers which flow in from the base electrode in order to retain space-charge neutrality, exceeds the normal ($I_e = 0$) density of majority carriers, so much so that the injection ratio falls and the life-time of the minority carriers falls, reducing the base transmission factor; $(1 - \alpha)$ therefore rises for two reasons. A more detailed study is required of these phenomena before we can say to what extent variation of α with I_e can be reduced.

Dr. J. Evans: In Fig. 4 of the paper by Messrs. Hilbourne and Jones there is the usual curve showing current gain decreasing with rise of emitter current. It would be better to make these power amplifiers with an *n–p–n* alloy transistor rather than a *p–n–p*. The authors mention that the difference in this curve of α_{cb}/I_e for the two types would introduce complications in complementary symmetry. If restricted to one type of transistor, I would use the *n–p–n*, because the published data seem to suggest that its current gain varies very much less with emitter current than that of the *p–n–p*. I presume that carbon tetrachloride is used because it is non-inflammable and can put out the fires caused by the transistor. It is now possible to get fluoro-carbons with very good properties for removing heat from transistors, and which are more inert than carbon tetrachloride.

In the paper by Mr. Light and Mrs. Hooker the only thing I understood was the fact that they require a transistor with a higher collector voltage. This could be obtained quite satisfactorily, since a high frequency is not required, by using a material of high resistivity. Normally, for a low base resistance it is necessary to use a low-resistivity material, but the authors' requirement is not so serious and it is possible to use a higher resistivity and get a higher collector voltage. For this application *p–n–i–p* would be ideal, and I wonder whether the authors have tried it.

* "Noise in Silicon Microwave Diodes," *Proceedings I.E.E.*, Paper No. 1671 R, September, 1954 (**101**, Part III, p. 317).

In Section 6 of his paper Mr. Stephenson states that he can find no dependence of noise on collector voltage, and that previous reports of a strong dependence were made on early transistors which were not hermetically sealed. I disagree. We have repeated this test on several transistors and definitely find a strong dependence of the noise on collector voltage, as we measure it. I do not know how this differs from the measurements made by Mr. Stephenson, but I suspect that the lack of dependence which his readings portray is a function not of the encapsulation but of the method by which the noise was measured.

Mr. E. H. Cooke-Yarborough: Since there is no current reversal in the transformers in the paper by Mr. Light and Mrs. Hooker, presumably less than half the flux curve is used. Have the authors considered biasing the flux in the transformer, perhaps with a permanent magnet, to allow the core to be used more efficiently?

We are extremely interested at Harwell in small portable convertors of the type described, and we use them a good deal in portable radiation monitors for supplying the high voltages needed by Geiger–Müller counters or photomultipliers. They must be able to operate over a wide range of battery voltage with little change of output voltage, because a nominal 4½-volt dry battery should be used until it has discharged to 3 volts. We do not like to carry a second battery to use as a voltage reference, and look forward to using the so-called "Zener" diodes when these become available. At present we favour the use of a parallel stabilizer consisting of a corona tube or an oscillating gas triode, shunted direct across the output terminals of a voltage-multiplying rectifier. This has the effect of limiting the amplitude of the square wave delivered by the oscillator to the rectifier system and allows stable voltages different from that of the reference to be obtained—either by adding further stages to the multiplying rectifier or by rectifying the square wave taken from another secondary winding on the transformer. (We obtain a stable 6-volt supply by this means.) The stabilizer also has the effect of limiting the voltage swing applied to the base of the oscillator, thus limiting the current drawn from the battery by the oscillator. Consequently, when the stabilizer is connected, the current drawn from the battery falls instead of rising, as might be expected from the paper. Stabilization by this means is therefore not unduly wasteful. We have a 1650-volt power supply for a photomultiplier which is stabilized in this way and which draws a constant 8 mA from a battery of any voltage between 3 and 4½ volts.

No discussion on transistors is complete without an argument about the drawing of transistor circuits. In drawing valve circuit diagrams, positive leads are usually drawn towards the top of the diagram and negative leads towards the bottom. Mr. Light and Mrs. Hooker quite often—though not always—put the negative parts of the circuit at the top. It is particularly necessary to be consistent about polarities in transistor circuits, because, with a valve, we do at least know that the anode conducts only when it is positive, but a transistor collector may conduct in either direction. Fig. 8 provides an example of the confusion which may be caused by non-adherence to a polarity convention. On the left positive is at the bottom and on the right it is at the top; since the circuits are interconnected, it is very difficult to follow what is happening.

Valve circuits are usually drawn positive upwards and waveforms on oscillographs are usually displayed positive upwards, so it is logical to draw transistor circuits with positive upwards. If the most positive potential happens to be earth, and the rest of the circuit is drawn underground, what does it matter? The only argument I have heard for the negative-upwards convention is that a *p–n–p* junction transistor behaves somewhat like a valve, and some people feel more at home seeing the electrode equivalent to the anode placed towards the top. Even this argument breaks down where *p–n–p* and *n–p–n* transistors are mixed, as in Fig. 13 of the paper by Messrs. Hilbourne and Jones, which would have been very confusing had a logical polarity convention not been observed.

Mr. J. W. R. Griffiths: Mr. Stephenson originally made the error of adding the individual contributions from the various sources on a voltage basis rather than a power basis—a method which would have been perfectly sound had the sources been completely correlated. The corrected results, in which the addition has been made on a power basis, assume no correlation between the sources. In fact, the true situation is somewhere in between. Three main individual phenomena contribute to the total noise in a transistor, namely semi-conductor noise, which is predominant at low frequencies, shot noise and thermal noise; hence the correlation between the two generators shown in Fig. 6 will depend on the contribution from each source, and will thus be a function of frequency. The curve (Fig. 7) relating F_{min} to R_s/R_{opt}, since it is dependent on correlation, is also dependent on frequency, and consequently it is of little practical value.

Experiments made at H.M. Underwater Detection Establishment on transistors (hermetically sealed transistors of the same type as those tested by Mr. Stephenson) have shown that the noise factor is markedly dependent on collector voltage when the latter exceeds a threshold value which varies from sample to sample of the same type of transistor. In this respect, Mr. Nicoll shows that semi-conductor noise is a function of current, so it would be more logical to represent the output noise as being produced by a current generator in the collector. The variation of noise factor with collector voltage might then be explained by a variation of the collector resistance, but, unfortunately, the magnitude of the variation of the latter is insufficient to explain the large changes in noise factor which occur.

Dr. G. B. B. Chaplin: The two applications papers have a realistic approach and do not seek to minimize the limitations of the junction transistor. This is refreshing and is more effective in establishing the junction transistor than some of the rather over-optimistic statements sometimes made.

In Section 9 Messrs. Hilbourne and Jones described the operation of a relay by a transistor. This is a very useful application, since the transistor will operate from the same supply as that required by the relay. For example, at the Atomic Energy Research Establishment we have operated an electro-mechanical register having a nominal resistance of 480 ohms. The transistors are connected in a cross-coupled monostable circuit, one of them supplying the 50 mA necessary to drive the register and the other fixing the duration of this current at 0·1 sec. The circuit is triggered by 1 microsec pulses. In the last sentence in Section 9 the authors say that the switching time of such a circuit is 2 millisec. Does this refer to the build-up time of the current in the relay inductance, or should it be 2 microsec referring to the rise time of the collector voltage?

In the footnote to Section 2.2.3 concerning the reduction of the hole-storage effect Mr. Light and Mrs. Hooker suggest that a positive pulse can be applied to the base during the switching-off transient from a tapping on the secondary winding. Fig. 1 shows that this pulse is already present on the base winding, and if a condenser is shunted across R_b it will be applied to the base with the desired effect. The condenser can be recharged by connecting a resistor from base to earth.

The use of a variable resistance in the same Figure to compensate for variations in current gain between transistors is undesirable. The resistance defines a certain base current, which, in turn, defines a collector current of $\alpha' I_b$, and the duration of the

"on" state is the time taken for the collector current to increase almost linearly to this level.

The operation of the circuit in Fig. 1, however, can be made less dependent on current gain by the base circuit shown in Fig. A. R_b is now much smaller and the base current is initially

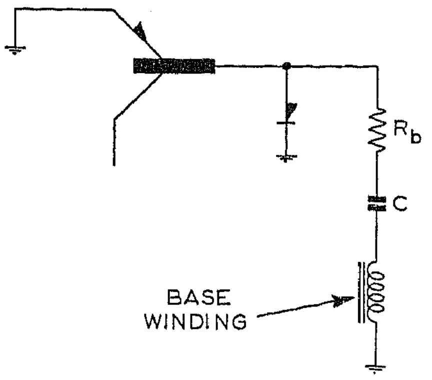


Fig. A.—Modified base circuit.

high but decreases exponentially as the condenser discharges, giving a more accurate intersection with the increasing collector current. Advantages which accrue from this modification are the possible elimination of the preset resistance, augmentation of the base current when switching on, the cutting off of emitter current as soon as the collector ceases to be current-saturated and automatic self-starting.

I endorse all that has been said concerning the very desirable positive-upwards polarity convention in circuit diagrams, but feel that it is closely connected with the choice of suitable transistor symbols. These symbols should clearly differentiate between point-contact and junction transistors and if possible bear some resemblance to the physical construction. The conventional symbol resembles the construction of the point-contact unit, and when turned on its side allows clear circuit diagrams to be drawn using the positive-upwards polarity convention. In 1953 a paper on point-contact trigger circuits* appeared using a symbol in which the emitter and the collector are on opposite sides of the base. This strongly resembles the construction of the junction transistor and could with advantage be adopted as the junction symbol.

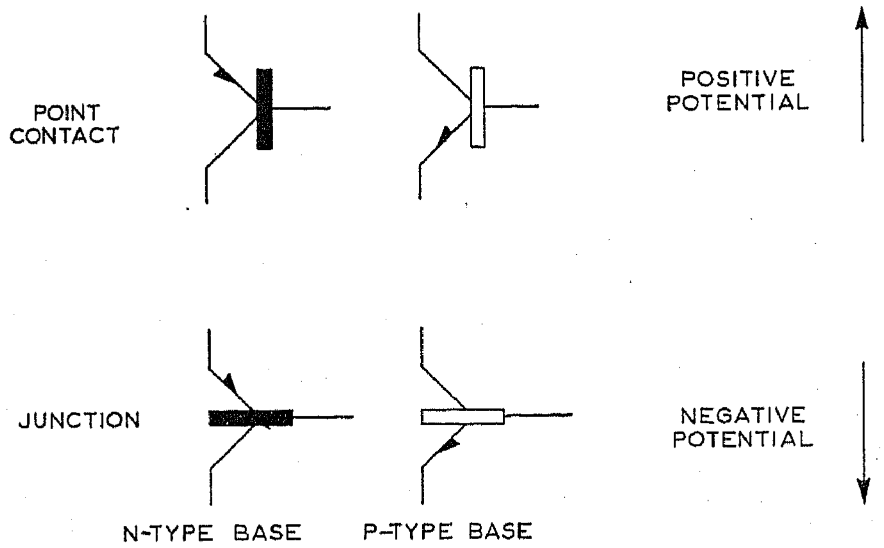


Fig. B.—The four main symbols.

It is also necessary to discriminate between *n*-type and *p*-type transistors, and this can be achieved by using a solid base to represent the *n*-type and an open one for the *p*-type. The resulting four main symbols are shown in Fig. B, together with the voltage convention. The arrows indicate the emitters and the symbols are readily extendible to include other configurations of *p–n* junctions.

* Williams, F. C., and Chaplin, G. B. B.: "A Method of Designing Transistor Trigger Circuits," *Proceedings I.E.E.*, Paper No. 1428 R, January, 1953 (100, Part III, p. 228).

Mr. J. N. Barry: There are two points I should like to raise on the paper by Messrs. Hilbourne and Jones.

The first concerns the relative merits of the common-collector and common-emitter circuit configurations under class-B push-pull operating conditions. The authors tend to favour the common-collector class-B push-pull amplifier. I think it depends very much on the practical application which type of arrangement is to be preferred, for the following reasons.

The authors show that under comparable conditions the common-collector arrangement has an inherently lower distortion than the common-emitter arrangement, but also a considerably lower power gain. They point out that this means that for a given output power it is necessary to have a higher driving power into the common-collector arrangement and also that, owing to the need for having a low driving-source resistance, it is also highly desirable that the driver itself is a common-collector type of circuit. This means that the driver itself is also acting as a power amplifier and may not be correctly matched.

I feel that in many practical cases, e.g. radio receivers, the loss of gain incurred by using common-collector driver and output stages may be a serious consideration. In addition, it is advisable to maintain matching between stages as far as possible, for otherwise the potential performance of transistors as amplifiers is being impaired appreciably. The distortion figures which the authors derive for the common-emitter arrangement refer to the use of the peak currents permissible in the transistors considered, but in many cases it may be possible to operate with lower current swings and hence obtain lower distortion. Is it possible that the recommendation made by the authors may also be modified in the future owing to improved performance of later types of transistor?

It seems noteworthy that all American transistor receivers which I have seen use common-emitter driver and output stages. Might it not be possible in many cases to obtain a better overall performance with an acceptable distortion level by using a common-emitter-connected output stage preceded by a common-emitter driver stage, together with the application of a certain amount of negative feedback between the output load and the input of the driver stage?

The second point refers to the experimental results presented in Table 2, where the authors derive a power efficiency of 69% from the practical measurements, and then, by taking into account the quiescent power drain, derive a higher efficiency figure of 76%. I think some qualifying remark should be made in applying this particular correction, for it is not of a general nature. If, for instance, the amplifier was operated under conditions much nearer class AB than class B the quiescent power drain would rise quickly, and "efficiencies" of the order of 100% or higher could theoretically be obtained.

Mr. E. Wolfendale: Transistors are made in which the current gain is maintained reasonably constant up to high currents. For example, with a small power transistor the current gain may be maintained fairly evenly up to a collector current of the order of 120 mA, and this is carried on in higher-power transistors up to 1 amp or more. This would probably cast a different light on the comparison of the earthed-emitter and earthed-collector output circuits.

Where power transistors are used in amplifiers we are often very much concerned with the standing current required for the amplifier. The smaller the current required for the class-A driving transistor the better for the life of the battery. This current is dependent on the gain in the output stage, and in many cases the earthed-emitter output stage is very important

owing to its increased power gain compared with the earthed collector.

Have Messrs. Hilbourne and Jones considered the possibility of using a tapped speaker in the output stage to avoid the use of the output transformer, so eliminating the other disadvantage of the earthed-emitter stage—the design of the output transformer.

Mr. F. Oakes: Can Messrs. Hilbourne and Jones say whether the extra gain provided by earthed-emitter amplifiers could be used in the form of negative feedback to improve the distortion figures, so that they will be better than those of the earthed-collector amplifier?

Dr. E. Billig: I was first led to consider thermal breakdown in rectifiers during a study of the reverse characteristics of selenium rectifiers some five years ago. The principles of thermal instability were fully discussed in two papers,* which gave, amongst others, the equivalents of eqns. (1)–(5) in the paper by Messrs. Waddell, Mayer and Kaye. It is gratifying to note that some of these early predictions, which seemed rather startling at the time, have now been verified by experiment. In particular, the prediction that what limits the rating of the rectifier—or, indeed, of any device with a negative temperature/resistance characteristic—is not so much the maximum voltage, current, or even the ambient temperature and power loss *per se*, but simply the temperature rise of the device above ambient temperature. Even the actual value of the critical rise reported here, 11° C, seems to agree very closely with the value to be expected, for instance, from Fig. 5 of the first Reference.

Mr. T. W. Sheppard: Section 2.1 of the paper by Messrs. Waddell, Mayer and Kaye is particularly interesting, since considerations of thermal stability apply to transistors as well as to diodes. There has recently been a paper† which discusses this subject in more detail. It is an intriguing fact that, no matter what the ambient temperature, the temperature rise above this before instability occurs will be about 11° C for germanium. However, this does not mean that the permissible reverse voltage will be the same at all ambient temperatures: it will have to be reduced as the ambient temperature is increased, because the leakage current also increases while the power dissipation for a given temperature rise remains the same.

Fig. 1 may be capable of some misinterpretation if care is not taken. The losses which are not temperature-dependent cause an internal temperature rise which is equivalent in its effect to a rise in ambient temperature. However, the critical temperature rise before instability occurs will still be 11° C, but now referred to this effective ambient temperature, so that the temperature rise above the actual ambient temperature is increased, as Fig. 1 shows. But the maximum permissible reverse voltage is still reduced below the value when no other losses are present, in the same way and for the same reason that it is reduced by an actual increase of ambient temperature. Moreover, since the temperature rise caused by the non-temperature-dependent losses depends on the thermal capacity of the unit, I consider that Fig. 1 is applicable only for a given thermal capacity and is not a general curve, as might at first be thought.

I think that some mention of thermal instability might have been made by Messrs. Hillbourne and Jones, particularly since it appears that in the basic circuits (Figs. 6, 8 and 9) there is no provision for preventing the collector leakage current increasing, owing to transistor temperature rise, and causing a further temperature rise because of the increased collector dissipation until thermal runaway occurs. It is interesting to estimate the chances of thermal instability occurring, and I have done this by assuming that the leakage current at a temperature T° C above 20° C is increased over its value at 20° C by the factor ε^{BT} where B is a constant whose measured values range from 0·07 to 0·1.

Taking the common-collector amplifier described in Section 6.3 I find that the temperature rise due to the 105 mW loss incurred at a supply voltage of 22½ volts would be 21° C, using the value of 5 mW/°C given in Section 3. If the actual ambient temperature was 20° C the effective ambient temperature would be 41° C. If the leakage current at 20° C is 5 mA, the resultant temperature rise above the effective ambient would be only about 1° C, and since 11° C is allowed, there is no danger at all of this circuit running away; but it is a different story if the ambient temperature is increased.

The actual temperature at which thermal instability will occur depends a good deal on the value of B, which can range between 0·07 and 0·1. For the most favourable case I calculate that stability is possible up to an effective ambient temperature of 78° C, but, for the worst case, only up to 51° C. If the power dissipation due to the non-temperature-dependent losses is kept at 105 mW, the maximum ambient temperatures would be only 67 and 30° C respectively, but if higher ambient temperatures are to be experienced this value of 105 mW must be reduced. However, at the most it can be reduced to zero, which means that 78° C and 51° C are the upper limits of ambient temperature. If still higher ambient temperatures are required, the collector voltage must be reduced or steps taken to increase the cooling of the transistor.

These comments on thermal instability should warn transistor users to avoid higher powers, temperatures or collector voltages than the makers recommend; but information often tends to be rather meagre, and if attempts are made to increase power output or use higher ambient temperatures, this is something which ought to be borne in mind for this sort of circuit, although not, of course, for all circuits.

Mr. F. F. Roberts: Measurements of noise figure with earthed base made at Dollis Hill on "quiet" junction transistors from several sources (cf. the paper by Mr. Stephenson), show a continual fall with increasing frequency from 75 c/s up to about 2 kc/s (with 1 volt and 1 mA at the collector), followed by a constant noise figure up to at least 9 kc/s; no significant subsequent increase of noise figure is expected until frequencies of 200–300 kc/s are exceeded. The low-frequency frequency-dependent noise figure is found to be sensitive to collector voltage in the range 1–5 volts. Other measurements on unprotected units exposed to controlled atmospheres indicate that similar voltage-dependent noise is associated with water vapour present in amounts insufficient to cause significant deterioration of other characteristics. The current-dependence of the noise figure in "quiet" units is admittedly more marked than the voltage-dependence, but the effect is again predominantly in the low-frequency frequency-dependent portion of the spectrum (which, for a 10 mA collector current, may extend up to above 9 kc/s).

In eqn. (1) of the paper by Messrs. Waddell, Mayer and Kaye W_i is given as the intrinsic energy gap. This is true only if the minority carrier life-time is independent of temperature. However, it is often found by independent measurements on the bulk material that the life-time increases exponentially with the negative reciprocal absolute temperature. One-half of the corresponding activation energy should be subtracted from the intrinsic energy gap to give the W_i used in eqn. (1).

The authors comment on the fact that the cut-off of sensitivity

* BILLIG, E.: "Application of the Image Force Model to the Theory of Contact Rectification and of Rectification Breakdown," *Proceedings of the Royal Society*, A, 1951, 207, p. 156.
BILLIG, E.: "Thermal Instability of Contact Rectifiers: the Effect of the Constituent Materials on the Efficiency of the Rectifying Junction," *Proceedings of the Physical Society*, B, 1951, 64, p. 342.
† BRIGHT, R. L.: "Junction Transistors used as Switches," *Communications and Electronics*, March, 1955.

on the short-wave side is steeper than 6 dB/octave. One possible explanation of that is surely that the absorption depth of the light in the germanium varies very steeply with wavelength, and from the published data it appears that for a wavelength less than 0·8 micron the absorption depth is less than 0·1 micron, which is so thin that it is probably dominated by the surface properties of the germanium, as, for example, the presence or absence of "channel." We carried out admittance/frequency measurements, at zero bias at room temperature without illumination, on a few of the authors' diodes some time ago and inferred carrier lifetimes of about 15 microsec.

Commenting on Messrs. Hillbourne and Jones's interpretation of Webster's explanation of α, I think that we should not be too ready to accept Webster's explanation that surface recombination is dominant. We have observed the properties of many junction transistors (without encapsulation) in controlled atmospheres and find that ambient atmosphere changes which cause considerable variations in surface recombination (as measured by separate experiments on bulk material) have very little effect on the current gain. Is it not possible that changes in the diffusion "constant" with carrier density might equally well explain the initial rise of gain with emitter current?

Mr. D. L. Johnston (*communicated*): The five practical voltage convertors described by Mr. Light and Mrs. Hooker are particularly interesting, and it would be useful to know the optimum operating frequencies in each case, as specified in Fig. 4. Presumably it is in the region of 10–20 kc/s for the smallest convertors and decreases as the power level increases.

Our own experience has been mainly with low power levels, typified by the first and third examples. The authors refer to this circuit as a class-C oscillator (Reference 5). In fact, this mode of operation is adjusted with sufficient feedback to bring about square-wave operation more akin to the mechanism of the authors' oscillator. The oscillator is non-critical and will work with almost any transistor without re-adjustment.

This suggests that low-grade transistors unacceptable for normal use may be employed as convertors; do the authors feel that the qualities of semi-conductors are such that a low-amplification transistor is likely to have average reliability, or is the low characteristic associated with inferior quality of the germanium?

A very thorough field trial of the small h.v. convertors has been undertaken, and about 5 000 units have been manufactured. It is a tribute to the standard that can now be achieved in transistor manufacture that failures have been negligible. Some 2 000 of these units have been made up in encapsulated form using epoxy resin. Production losses and failures in service have been well below 1%, and this points to the potted "brick" as a very promising mechanical construction in transistor circuits. After encapsulation the unit is virtually indestructible, and this overcomes the objection that might be levelled at the very small components and coil employed in the small unit shown in Fig. 13.

THE AUTHORS' REPLIES TO THE ABOVE DISCUSSION

Mr. G. R. Nicoll (*in reply*): Dr. Tillman has raised the complicated problem of the barrier height and its variations. This problem has been avoided in the paper by postulating a model of just sufficient complexity to correspond with the results of the noise measurements.

The essential feature used in deriving the magnitude of shot noise, eqn. (11), is that the current through the diode consists of two uncorrelated components which flow in opposing directions and which are in the ratio $\exp(eV/kT):1$. Eqn. (11) thus contains only the quantities i, V and T, which can be measured direct, and the barrier height does not appear explicitly. (Attention has been drawn to a numerical error in eqn. (12) which should read

$$\overline{i^2}/4kT = t/R = 20i\coth(20V)$$

The correct form of the equation was, in fact, used in the analysis of the measurements).

Similarly, the barrier height has not been used numerically in treating flicker noise, but is introduced to provide a simple picture which agrees with the general results of the measurements.

As Dr. Tillman indicates, it is not easy to determine satisfactory values for the barrier height: this applies equally to silicon and germanium point-contact diodes. The physical difficulties of the simple model should perhaps be stressed; a more satisfactory model may ultimately be provided within the framework of transistor physics.

Mr. W. L. Stephenson (*in reply*): Dr. Evans has expressed disbelief in the lack of variation of noise with collector voltage, and distrusts the method of measurement. I should like to point out that by the same method measurements on transistors encapsulated in plastic agree with published figures.* However, the results given in the paper apply to transistors hermetically sealed in glass.

The maximum collector voltage used in these measurements was 6 volts, as shown in Fig. 11. That the noise factor may vary with voltage above this level has not been considered, since this voltage was considered sufficiently great for amplifiers handling signal levels at which the noise factor is important. It can be expected that the noise will increase as the collector voltage approaches the turnover point.

I am indebted to Mr. J. W. R. Griffiths for originally pointing out the error in Section 4.3 by which the noise contributions were added on a voltage rather than a power basis, thereby assuming complete correlation between the sources.

In correction of this, the individual voltages may be considered as $K\sqrt{(R_s)}\cos\omega t$, $V_g\cos(\omega t+\theta)$, $I_g\cos(\omega t+\phi)$, so that $\psi=(\theta-\phi)$ is the correlation angle between V_g and I_g.

The total instantaneous input voltage is therefore

$$\frac{R_{in}}{R_{in}+R_s}[V_g\cos(\omega t+\theta)+K\sqrt{(R_s)}\cos\omega t+I_gR_s\cos(\omega t+\phi)]$$

so that the total mean input power is

$$\frac{R_{in}}{2\pi(R_{in}+R_s)^2}\int_0^{2\pi}[V_g\cos(\omega t+\theta)+K\sqrt{(R_s)}\cos\omega t + I_gR_s\cos(\omega t+\phi)]^2d(\omega t)$$

$$=\frac{R_{in}}{(R_{in}+R_s)^2}\left[\frac{V_g^2}{2}+\frac{K^2R_s}{2}+\frac{I_g^2R_s^2}{2}+KV_g\sqrt{(R_s)}\cos\theta + I_gKR_s^{3/2}\cos\phi+V_gI_gR_s\cos(\theta-\phi)\right]$$

The contribution from R_s is

$$\frac{R_{in}}{(R_{in}+R_s)^2}\frac{K^2R_s}{2}$$

so that

$$F=1+\frac{V_g^2}{K^2R_s}+\frac{I_g^2R_s}{K^2}+\frac{2V_g\cos\theta}{K\sqrt{R_s}} + \frac{2I_g\sqrt{(R_s)}\cos\phi}{K}+\frac{2V_gI_g}{K^2}\cos\psi$$

* See Shea, R. F.: "Principles of Transistor Circuits" (Wiley, 1953), p. 441.

Since there cannot be any correlation between the noise in R_s and the transistor noise, θ and ϕ are random phase-angles so that their mean value is zero.

The value of F is therefore

$$F = 1 + \frac{V_g^2}{K^2 R_s} + \frac{I_g^2 R_s}{K^2} + \frac{2V_g I_g}{K^2}\cos\psi$$

Again, putting $R_s = n\frac{V_g}{I_g}$,

$$F = 1 + \frac{V_g I_g}{nK^2} + \frac{nV_g I_g}{K^2} + \frac{2V_g I_g}{K^2}\cos\psi$$

$$= 1 + \frac{V_g I_g}{K^2}\left(n + \frac{1}{n} + 2\cos\psi\right)$$

so that, as before, $$R_{opt} = \frac{V_g}{I_g}$$

but now $$F_{min} = 1 + \frac{2V_g I_g}{K^2}(1 + \cos\psi)$$

so that $$\frac{F-1}{F_{min}-1} = \frac{n^2 + 2n\cos\psi + 1}{2n(1+\cos\psi)}$$

where $n = R_s/R_{opt}$.

This expression replaces the graph of Fig. 7, which, as it stands, is of little practical use, but the formula given above which gives the variation of F with F_{min} and R_s/R_{opt} can be of use provided that the correlation angle ψ (which may be a function of frequency) is known.

Messrs. J. M. Waddell, S. E. Mayer and S. Kaye (*in reply*): In reply to Dr. Tillman, we are quite satisfied with the results obtained using a soldered seal in production. Special precautions are necessary, but current samples have satisfactorily withstood 84 days of tropical exposure to RCS 11.

In the noise measurements quoted the load used was 500 kilohms so that the lower values of noise are admissible. However, not all cells show this index; some have lower values, the lowest observed being about $-1\cdot25$.

When the data for Fig. 5 were obtained the beam of incident light had a half-angle of the order of arc sin 1/6, so that the cone in the germanium would have a half-angle of the order of arc sin 1/24, and after travelling 800 microns would have diverged by about 33 microns. In an endeavour to avoid this type of error the spot was focused approximately 200 microns away from the junction. The type of mechanism first given and also any internal scattering would give rise to a discontinuity in the curve for the complete cell.

We agree with Dr. Billig that the theoretical predictions of thermal instability appear rather startling at first sight; the development of practical devices having the appropriate theoretical characteristics has enabled the theory to be verified experimentally and has at the same time emphasized its practical importance.

We do not altogether agree with Mr. Sheppard that the theoretical results expressed in Fig. 1 can be derived by assuming an 11° C rise over an ambient temperature which has been increased by P_2/m; a simple numerical example will show this. If we take a value of $P_1/(P_1 + P_2)$ of $0\cdot2$, the simpler view would lead to a maximum rise of approximately 55° C, whilst the exact calculation gives a value of approximately 80° C. We do agree, however, that the maximum permissible reverse voltage will be reduced by the presence of additional losses. The theory given in the paper is limited to steady-state conditions and does not include the transient conditions on first switching on the reverse voltage. The thermal capacity of the device does not therefore appear in the equations, and the results given in Fig. 1 are independent of the thermal capacity and the value of m.

In reply to Mr. Roberts we have measured the reverse-current/temperature characteristics for these junctions, and they fit eqn. (1) if values for W_i are assumed which are in agreement (within the experimental error) with the commonly accepted value for the intrinsic energy gap of $0\cdot72$ eV. We agree with the explanation he gives for the steepness of the short-wave cut-off.

Messrs. R. A. Hilbourne and **D. D. Jones** (*in reply*): We should like to point out to Dr. Tillman, Dr. Evans and Mr. Roberts that the paper is not primarily concerned with the physics of transistors; however, we agree that a sound knowledge of the factors governing the shape of the α–I_e curve is very important. Unfortunately, the exact mechanism of this variation is not clearly understood at present. Webster in his explanation does not take into account the effect of the geometry of the transistor on the form of this curve. The shape of the emitter bead and its proximity to the base-contact ring appear to play an important part in the value of α at high emitter currents. The low value of α at low emitter currents is not so important in power amplifiers, since it is partially masked by the higher input resistance which exists under these conditions. The value of α at these low emitter currents is found to be closely connected with the surface treatment, in some units having a very flat peak to the α–I_e curve. It is also found that the peak value of α decreases if the device is operated for long periods at junction temperatures of the order of 70° C.

The figure quoted, of 1° C junction temperature rise per 5 mW of collector dissipation, refers to an experimental type of transistor. This is hermetically sealed in a metal can which is isolated from the transistor, the can being gas-filled. It was found that the temperature rise for a given dissipation is very dependent upon the lead wire thickness, which indicates that a large part of the cooling is due to conduction through these leads. The temperature rise is considerably reduced when the can is connected to one of the transistor electrodes. In reply to Dr. Tillman we would therefore suggest that the connection of a heat sink to the isolated can would not greatly reduce the rise in junction temperature.

Mr. Barry suggests that in some applications the common-emitter arrangement would be more advantageous than the common-collector arrangement. This is particularly true in applications such as radio receivers, in which the minimum possible number of transistors, consistent with a given performance, is an essential feature. Operating at a lower current level reduces the harmonic distortion obtained with a common-emitter amplifier, and in some cases this distortion may be less than the acceptable limit. However, if a low supply voltage such as a car battery is used, a high current level is essential to maintain the same output power level, and it may then be necessary to use the common-collector arrangement. The object of the paper was to indicate the limitations imposed by harmonic distortion on the three basic circuit configurations, and to show which of the transistor parameters are responsible for these effects. These limitations must be taken into account, together with the other important criteria, when designing power-amplifier equipment.

With regard to Mr. Oakes's question, it is important to note that a common-collector amplifier fed from a high-impedance source has virtually the same level of harmonic distortion as the comparable common-emitter amplifier. The slight difference in distortion levels is due to the small difference in the current gains of the two connections $1/(1 - d)$ and $\alpha/(1 - \alpha)$. It is not until the source impedance is reduced and negative-feedback effects occur, that the distortion level is decreased. If negative feedback is applied to a common-emitter amplifier such that the

distortion level is reduced to the value obtained with a common-collector stage, the power gains of the two stages are comparable. The common-emitter amplifier, with negative feedback, has the advantage that the amount of gain and distortion obtained can readily be varied.

Mr. Sheppard raised the question of temperature instability in junction-transistor amplifiers. This effect has been observed when using high-gain transistors at high values of collector voltage. Messrs. Waddell, Mayer and Kaye, in their paper on photo-electric cells, obtain the value of 11° C for maximum permitted rise in junction temperature, ΔT, by inserting values into the theoretical equation for the variation of reverse current with temperature. Experimentally the reverse collector current of a junction transistor is found to obey approximately the relationship

$$i_{co} = i_0 \varepsilon^{BT}$$

given by Mr. Sheppard. It can be shown that, with a device obeying such a law, the temperature rise necessary to cause thermal instability is given by

$$\Delta T = \frac{1}{B} \,{}^\circ\mathrm{C}$$

Since B has values between 0·07 and 0·1, ΔT can have values between 14° C and 10° C. This variation in the value of ΔT with B, which has been confirmed in practice, must also be taken into account when calculating the thermal stability of an amplifier at elevated ambient temperatures.

The power dissipated in the collector junction of a transistor due to temperature-dependent current varies with the circuit configuration. In the common-base connection, the power dissipated is given by

$$P_r = i_{co} V_c$$

which is low even at high temperatures and collector voltages. In the common-emitter and common-collector arrangements the expression for the reverse power becomes

$$P_r = \frac{1}{1 - \alpha} i_{co} V_c$$

With high-gain transistors operating with high values of collector voltage, the temperature-dependent power dissipation can become greater than the value required to produce thermal instability.

This power dissipation, and the resultant rise in junction temperature, can be reduced by applying d.c. negative feedback and hence reducing the effective value of $1/(1 - \alpha)$. In the common-collector amplifier a small amount of d.c. negative feedback is produced by the resistance of the transformer primary winding, when the bias voltage is obtained from a low-resistance source.

Mr. L. H. Light and **Mrs. Prudence M. Hooker** (*in reply*): The use of somewhat higher resistivity material as suggested by Dr. Evans would be advantageous in giving higher collector working voltages, but any large increase in base resistance is unwelcome because it increases both the base input power required and the transient losses. Transistor characteristics of the *p–n–i–p* type are indeed excellent not only for d.c. convertors but also for all forms of power switching.

We have experimented with magnetic bias, as mentioned by Mr. Cooke-Yarborough, with some success. The new ceramic permanent-magnet materials lend themselves well to pre-biasing ferrite cores, and with this combination an improvement in current-handling capacity of up to three times can be achieved using cores of non-uniform cross-section. With standard cores the improvement in core utilization is limited to some 1·5 times, which is worth while only if the utmost reduction in weight and size is required.

The stabilization method developed at Harwell is elegant and effective. It should be noted, however, that it is an intrinsic part of its action that the transistor is not normally bottomed, which reduces its power-handling capacity greatly. The method is thus best suited to low-power applications.

On the question of circuit-diagram layout, we too favour the practice of drawing the positive line uppermost because of its correspondence with oscillograms. On the other hand, the existing convention of having the earthy line at the bottom, as used in the Figures of the paper, does allow engineers with valve experience to gain more readily a rough idea of the functioning of many *p–n–p* transistor circuits, which is of value during their period of initiation into the mysteries of the transistor. A good opportunity for changing to the practice recommended by Mr. Cooke-Yarborough will arise when *n–p–n* transistors become more prominent.

The method of connecting across R_b the capacitor referred to by Dr. Chaplin is indeed the most generally useful one. R_b itself suffices to recharge the capacitor.

The circuit modification of Fig. A does not reduce the dependence of output on transistor characteristics sufficiently to allow the preset resistance to be omitted in view of the wide spread of α' existing in all types. Although this modification may ease starting in some cases, it does not make it automatic in all. It has the disadvantages of reducing the conversion efficiency and increasing the cost.

If it is a principal requirement that output voltage be independent of transistor characteristic it is advisable to use a type of convertor circuit in which transformer action and not ringing-choke action is used.* Such circuits have a low output resistance, but tend to be more complex, require larger transformers and have their efficiency greatly dependent on transistor characteristics.

Mr. Johnston's experience on transistor reliability is encouraging. We have no evidence to suggest that rejects for low α' are in any way less reliable than normal transistors, but if low-gain transistors were the result of an inferior manufacturing process they might well prove less stable. Apart from a slight drop in efficiency, due to their increased drive-power requirement, low-gain transistors should prove satisfactory.

The optimum frequency is a function of the size/power ratio; it is high (of the order of 10 kc/s) for the miniature unit (*a*), and of the order of 3 kc/s, 2 kc/s, 700 c/s and 1 kc/s (approximately) for units (*b*), (*c*), (*d*), and (*e*) respectively.

* UCHRIN, G. C., and TAYLOR, W. O.: "A New Self-Excited Square-Wave Transistor Power Oscillator," *Proceedings of the Institute of Radio Engineers*, 1955, 43, p. 99.

621.396.11: 621.317.34

Paper No. 1932 R.
Nov. 1955

AUTOMATIC RECORDING OF THE DIRECTION OF ARRIVAL OF RADIO WAVES REFLECTED FROM THE IONOSPHERE

By J. A. THOMAS, B.Sc., A.Inst.P., and R. W. E. McNICOL, M.Sc., A.Inst.P., Graduate.

(*The paper was first received* 16*th March, and in revised form* 4*th July*, 1955.)

SUMMARY

Equipment is described which, for a fixed frequency, simultaneously determines the direction of arrival of all resolved pulses reflected from the ionosphere by measuring the phase differences for two pairs of fixed orthogonal spaced-loop aerials. The vector sum and difference of the signals for each pair of loops in turn are formed in a special unit. These are amplified in normal receivers and the rectified video outputs applied to the Y- and X-plates respectively of a cathode-ray tube. The orientations of the resulting lines are thus governed by the phase differences between the signals at the pairs of loops. Pulse resolution is achieved by the simultaneous application of a time-base in the X-direction. Phase-balanced receivers are not required; since the phase-sensitive portion of the equipment is confined to the loops, pre-amplifiers and the sum-and-difference unit, the system is stable enough to run without attention for up to one week.

LIST OF SYMBOLS

α = Azimuth angle of downcoming radio waves.
δ = Elevation angle of downcoming radio waves.
ϕ = Phase difference between incident waves on two different aerials.
ϕ_1 = Phase difference between incident waves at N and S aerials.
ϕ_2 = Phase difference between incident waves at E and W aerials.
λ = Wavelength of radio waves.
d = Separation between opposite aerials of a pair.
$\Delta\phi$ = Phase difference corresponding to 1° change in direction of arrival.
$\Delta\phi(\alpha)$ = Phase difference corresponding to 1° change in azimuth.
$\Delta\phi(\delta)$ = Phase difference corresponding to 1° change in elevation.
R' = (Slant) virtual range = $\frac{1}{2}$ group path.
A_1, A_2 = Signal amplitudes at two spaced aerials.
S = Vector sum of A_1 and A_2.
D = Vector difference of A_1 and A_2.
$\theta/2$ = Angle of inclination of echo trace to the sum or Y-direction.
ρ = Amplitude correlation coefficient for radio waves received at spaced aerials.
$\bar{A}$ = Mean signal amplitude in all aerials.
$\overline{|x|}$ = Mean deviation of signal amplitude differences from the expected value.
$\overline{|\phi|}$ = Mean deviation of signal phase differences from the expected value.
b = Specular to non-specular amplitude ratio in a reflected signal.

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
Mr. Thomas and Mr. McNicol are in the Physics Department, University of Queensland.

(1) INTRODUCTION

For some years night-time records have been made at Brisbane and other stations in south-east Queensland of the virtual range of ionospheric reflecting points, as a function of time, at a fixed frequency of 2·28 Mc/s. At each station the receiver and pulse transmitter were housed together, and both used simple horizontal dipole aerials. The records obtained suggested that the positions of the reflecting points often departed considerably from the zenith; furthermore, comparison of records made simultaneously at three stations 100 km apart often indicated some horizontal motion of the reflecting points.

It was clear that interpretation of the records would be facilitated if data on the direction of arrival of the reflected pulses were available. In particular, a measuring system was required which would record (for all the echoes simultaneously) the slowly changing directions of arrival which were expected to be associated with reflections of the types shown in Figs. 1 and 2. Since the exact times of occurrence of such phenomena were unpredictable, it was also desirable that the equipment should work automatically.

Various workers have devised systems for measuring the direction of arrival of radio waves, and the most suitable of these for the purpose in hand was that of Ross, Bramley and Ashwell.[1] Since it was not desired to record the rapid diversity fluctuations

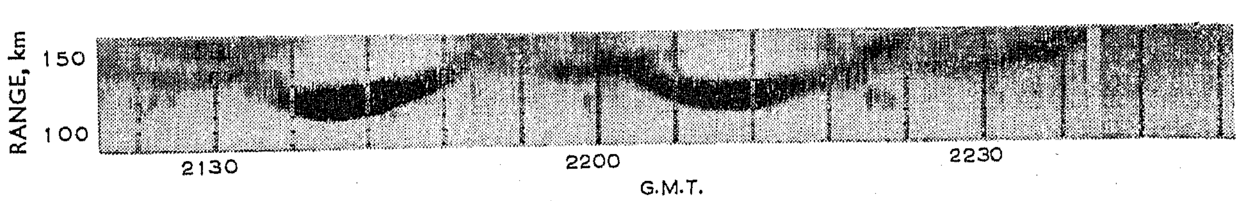


Fig. 1.—Virtual-path/time record for the 5th April, 1953, showing apparent motion of Es "clouds."
Transmitter frequency 2·28 Mc/s; pulse power output 1 kW; pulse length 70 microsec.

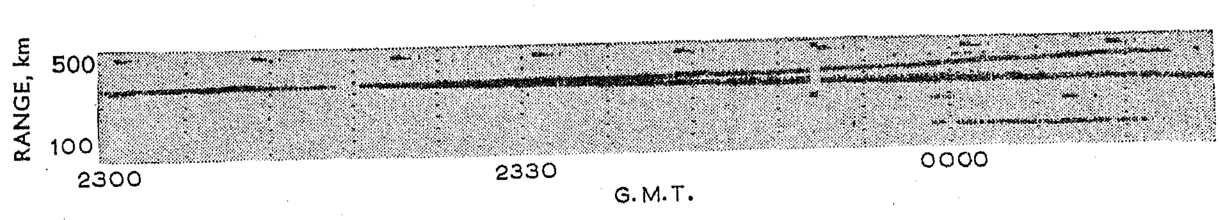


Fig. 2.—Virtual-path/time record for the 1st September, 1955, showing apparent motion of an F-region irregularity.

which were among the phenomena described by these authors, it was possible to modify their technique in such a way as to make the equipment capable of recording automatically the directions of arrival of all the resolved echoes. The sum-and-difference technique was partially retained; signals proportional to the vector sum and difference of the signals induced in two spaced aerials were amplified in separate receivers, and the rectified outputs from these receivers were displayed on a cathode-ray tube in such a way as to compare the magnitude of the sum signal with that of the difference signal. While this procedure caused the loss of some of the information in the original signals, it had the compensating advantage that, since the main receivers were required to be matched in gain only, the system was stable for periods of approximately one week. Application of a time-base synchronized with the transmitter pulses separated out the various echoes in the display. The resulting pattern was photographed with successive 3-min exposures; the integrating effect of this was such as largely to recover the accuracy lost by abandonment of the phase information mentioned above.

(2) GENERAL PRINCIPLES OF THE EXPERIMENTAL METHOD

(2.1) Basic Principles

A 4-aerial system was used, as indicated in Fig. 3. For a plane radio wave received at an azimuth angle α the phase advances of the electromagnetic fields are given by

$$\phi_1 = \frac{2\pi d}{\lambda} \cos\alpha \cos\delta$$

$$\phi_2 = \frac{2\pi d}{\lambda} \sin\alpha \cos\delta$$

If the four aerials are identical and have no mutual reactions, the phase difference between the r.f. signals from the N and S aerials

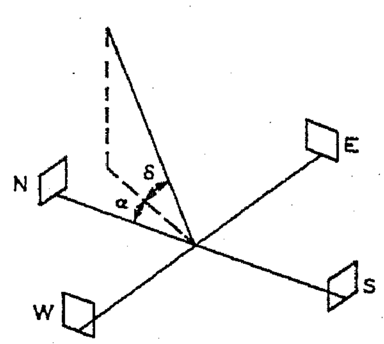


Fig. 3.—General layout of aerial system.

will be ϕ_1 and that between the E and W aerials will be ϕ_2. The angles α and δ may therefore be deduced from measured values of ϕ_1 and ϕ_2.

(2.2) Choice of Aerial System and Basic Limitations to Accuracy

The choice of a suitable aerial system for this type of measurement has been discussed by many workers.[1-4] Following Ross, Bramley and Ashwell, the system used consisted of two pairs of identical screened loops, mounted (as indicated in Fig. 3) so that opposite loops of a pair were parallel to one another and perpendicular to the line joining them. Such a system has the advantage of freedom from polarization and mutual-coupling errors, and also some reduction in site errors compared with vertical aerials.

Polarization errors occur when the responses of the two aerials of a pair are not identical for waves of all polarizations incident in any direction. This type of error is reduced to a minimum by mounting the two aerials of a pair parallel to within $\frac{1}{4}$°, and perpendicular to their feeder lines which run along the line joining them. Electrical unbalance ("antenna effect") is almost entirely eliminated by using double-turn screened-loop aerials with the gap in the screen accurately located in the centre of the lower horizontal limb.

As pointed out by Ross, Bramley and Ashwell, the fact that the phase angles ϕ_1 and ϕ_2 are measured for differently polarized components leads to no difficulties in practice.

The mutual impedance between the opposite aerials of a pair is considered to be negligible at the spacing used, namely $\lambda/\sqrt{2}$.

Site errors may be divided into "near" and "distant" errors, the near region covering the area of the site within a radius of a few wavelengths and the distant region that up to a radius of several kilometres. For the aerial system used in the experiment the errors due to reradiation from reflectors in both of these regions are negligible for directions of arrival near the vertical, and gradually increase with decreasing elevation angle.[1,5,6] The majority of all measurements made were for elevation angles greater than 45° and the distant site errors were ignored.

(2.3) Phase Measurement

As shown in Section 9.1, the ratio of the magnitude of the vector difference of the two r.f. signals induced in a pair of loops by a given echo to the magnitude of their vector sum is tan $\frac{1}{2}\phi$. Voltage pulses proportional to the vector sum and difference were applied to the *Y*- and *X*-plates of a cathode-ray tube, giving rise to a straight line at an angle $\frac{1}{2}\phi$ to the *Y*-direction. Diversity effects produce a flicker in the orientation of the line (see Section 9.2), and photographic integration was used to obtain the mean orientation over a period of about 3 min.

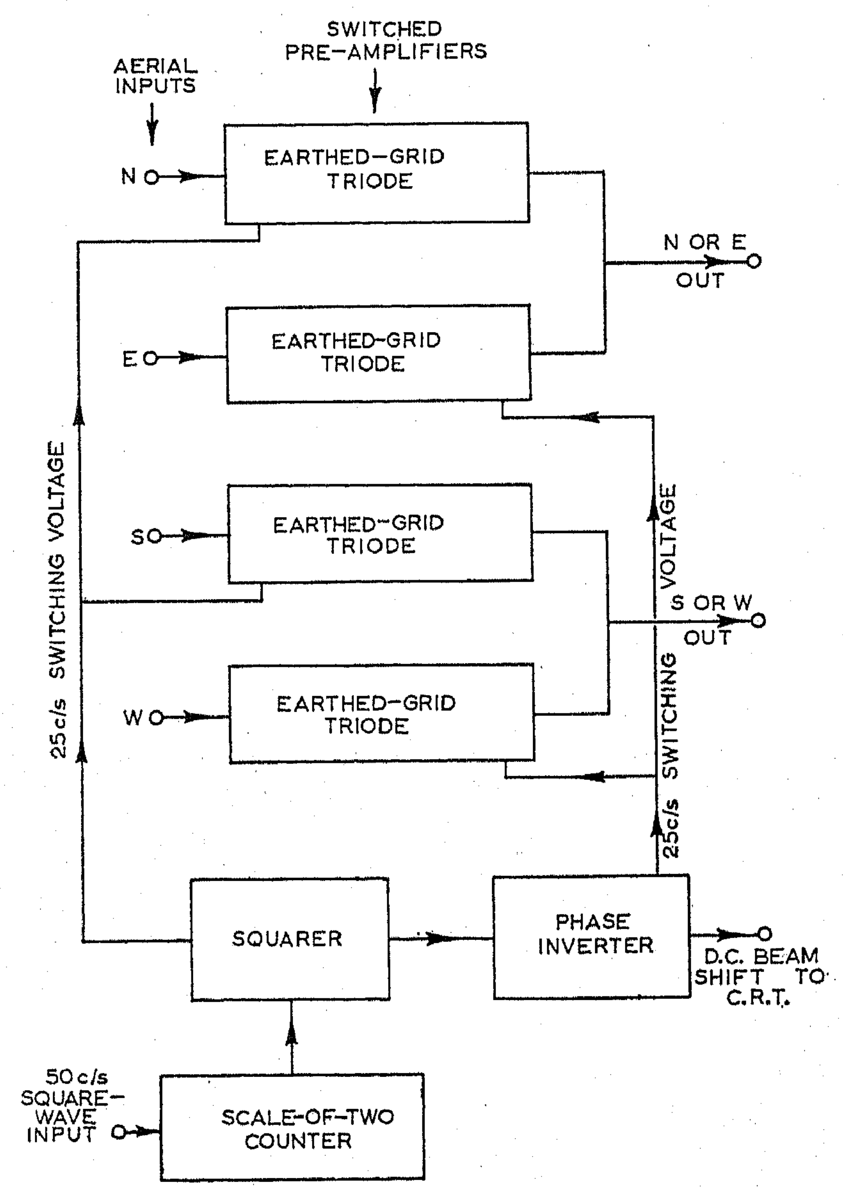


Fig. 4.—Schematic of electronic switch.

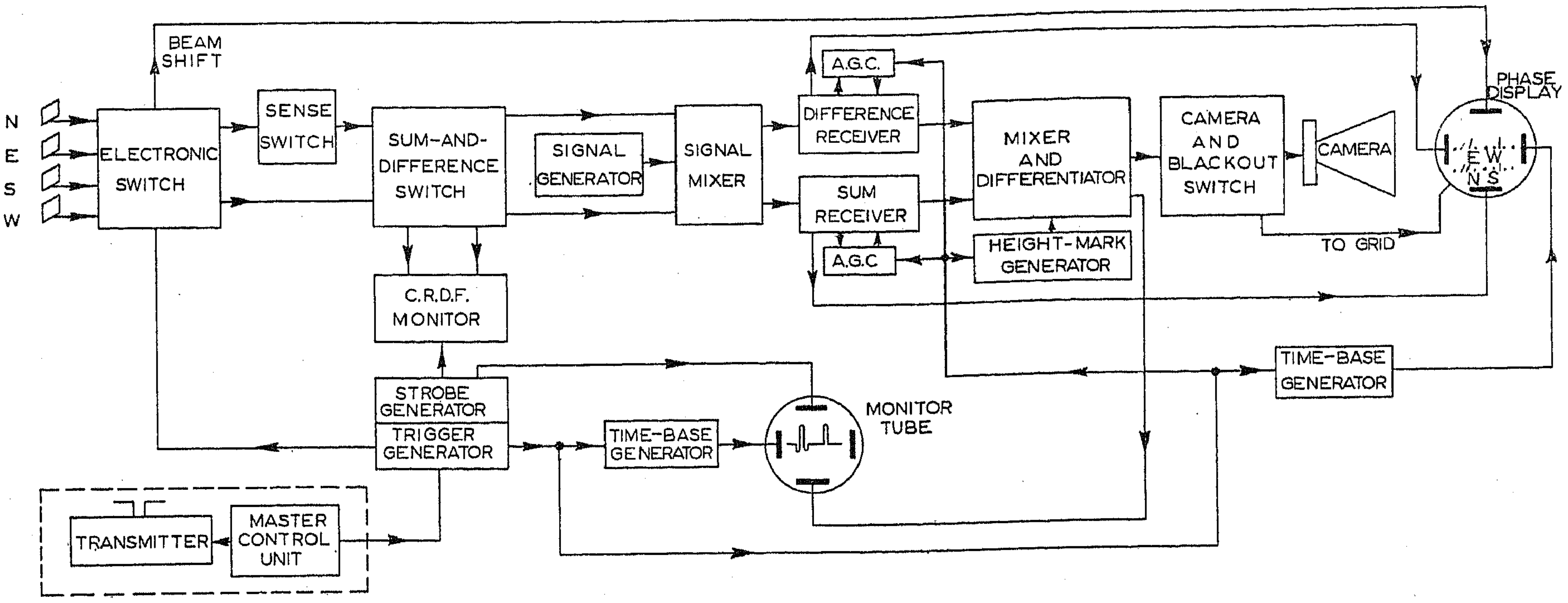


Fig. 5.—Schematic of complete equipment.

For the determination of sense a phase-shifting network was inserted in one of the leads between the electronic switch and the sum-and-difference unit.

(3) DETAILS OF AERIALS AND ASSOCIATED SWITCHING

The screened-loop aerials are attached to 6 ft square wooden frames and mounted with the lower limit 2 ft above ground level in such a way as to permit alignment of opposite pairs. This was achieved to an accuracy better than $\frac{1}{4}°$ with the aid of a theodolite. The spacing between the aerials in each pair was $\lambda/\sqrt{2}$.

The four aerials were transformed-coupled to coaxial feeder cables of 50 ohms impedance. The feeder cables were led into the central receiving hut in trenches 3 ft deep.

The mains and telephone cables were led into the receiving hut in a trench 3 ft deep running along the line of one aerial pair.[1] All other conductors were kept as remote as possible, in order to minimize errors due to reradiation. For site reasons the aerials were actually aligned 20° and 110° east of north instead of due north–south and east–west.

The two pairs of r.f. signals from the loops were selected alternately by an electronic switch (shown schematically in Fig. 4), which was synchronized with the transmitter. Successive ground pulses with their associated trains of echoes were thus passed alternately from the N–S and E–W pairs of aerials to the sum-and-difference unit. The electronic switch also provided a beam-switching voltage to displace the cathode-ray-tube trace vertically in synchronism with the aerial switching. The upper trace on the display was associated with the E–W pair of aerials and the lower with the N–S pair.

(4) DETAILS OF THE PHASE-MEASURING EQUIPMENT

The general method of achieving the phase measurement has been referred to in Section 2.3 and a schematic of the complete equipment is given in Fig. 5.

(4.1) The Sum-and-Difference Unit

The outputs from the aerial switching unit were connected to inputs of the sum-and-difference unit (Fig. 6). The first stage in this unit consisted of a pair of earthed-grid triode amplifiers, the split output coils of which were arranged to be in phase in one case and in antiphase in the other. (Exact balance of the amplitudes of these outputs was easily achieved by slight relative movement of the coils.) These outputs were fed to two earthed-grid pentode mixer stages, each consisting of a pair of pentodes with their anodes in parallel. The inputs to the mixers were so connected that one gave an output proportional to the vector sum and the other an output proportional to the vector difference of the signal voltages appearing at inputs. The gains of all four pentodes could be varied by adjustment of their

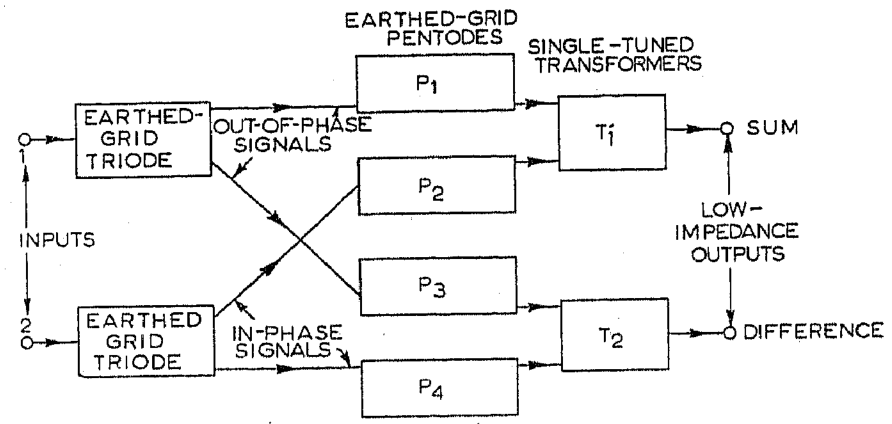


Fig. 6.—Schematic of sum-and-difference unit.

screen voltages; this compensated for differences in valve transconductance without introducing phase changes in the process. The sum and difference outputs were fed separately to the two gain-matched receivers.

Once the vector sum and difference of the signal voltages have been formed, there is no further need to preserve the phase of the signals, for reasons explained in Section 9.1.

In order to determine the sign of the phase difference between the signals in the pairs of aerials, at intervals during recording a small transformer (the input side of which was series tuned to resonance) was switched into the lead between the N and E outputs from the aerial switch and input[1] of the sum-and-difference unit. The transformer turns ratio was adjusted until the input and output signal voltages were equal. Insertion of the transformer had the effect of advancing the phase of the signals from the N and E aerials by 90°.

(4.2) The Gain-Matched Receivers

The receivers were of the Loran (AN/APN4) type,[7] which had been modified to prevent long paralysis after the transmitter

pulses. The two receivers were adjusted to have identical response curves by means of a swept-frequency alignment generator with cathode-ray-tube display.

It was essential that the ratio of the gains of the two receivers, once set to the correct value (determined by the *X*- and *Y*-sensitivities of the cathode-ray tube), should remain constant. Accordingly provision was made for automatic gain stabilization of the individual receivers. A special control pulse, obtained from a pulse-modulated signal generator, was fed at equal amplitude to both receivers in such a way as not to interfere with the normal connection to the receivers of the outputs from the sum-and-difference unit. This control pulse was "gated out" from the receiver video outputs, rectified, smoothed and arranged to control the gains of the receivers. The gains of the receivers thus adjusted themselves individually to keep the video amplitudes of the amplified control pulses constant at the predetermined values, and hence the ratio of the receiver gains remained constant. The control pulse was generated about 300 microsec after the transmitter pulse; ionospheric echoes are never received at this time.

It was found that the inherent gain stability of the receivers was such that the automatic stabilization could be omitted. The special signal-generator pulse was nevertheless retained, since it provided a continual check on the correctness of the receiver gain settings. On every record the pulse appeared as a line at 45° to the vertical; if the ratio of the receiver gains were to change the orientation of this line would change, and from the new orientation the necessary correction to the readings could easily be found.

(4.3) The Display Unit

In the display unit the video output of the sum receiver was applied to one of the *Y*-plates of a cathode-ray tube; to the other *Y*-plate a beam-switching voltage derived from the electronic aerial switch was applied. The *X*-plates were fed with both the video output of the difference receiver and the output of a synchronized linear time-base generator.

The video outputs from both receivers were mixed with 50 km range marks, amplified, differentiated and used to produce brightness modulation of the cathode-ray tube, so that the trace was visible only during the rising portion of each echo or range mark.

The time-base generator was of a standard Miller type,[8] but was modified by applying to the grid of the run-down valve the same waveform (at a suitable amplitude), which was used to brighten the trace. This arrested the uniform motion of the spot across the screen for as long as the trace was being brightened, and thus eliminated a small error in the angular position of the line which would otherwise have occurred.

(4.4) Photography

The recording camera operated at $f/4 \cdot 5$ and used 35 mm film, each frame representing an exposure of 160 sec. The film advance between exposures took 20 sec and was effected by a small electric motor; during the film advance the cathode-ray-tube trace was blacked out completely. Every fourth exposure was a "sense run," i.e. the phase-shift network described in Section 4.1 was switched into circuit. The necessary switching was performed by a mechanical switch driven by a continuously running synchronous motor.

(5) OPERATION

(5.1) Lining Up

To line up the equipment a portable battery-operated squegging transmitter with a short vertical rod aerial was used. To check the E and W loop aerials this transmitter was placed on the line joining the N and S aerials. A modified twin-channel cathode-ray direction-finder receiver[9] was used to check the equality of the phase and amplitude adjustments of the loops themselves and of the two following stages, i.e. up to the inputs of the sum-and-difference mixer stages (P_1–P_4 in Fig. 6).

When these adjustments had been completed, input 2 of the sum-and-difference unit was removed, and the sum-and-difference outputs from P_1 and P_3 were adjusted for amplitude and phase equality. Amplitude balance was achieved by varying the screen voltages of P_1 and P_3 and phase balance by adjusting the tuning of the output transformers T_1 and T_2. Both inputs were then restored, thus providing equal-amplitude out-of-phase signals at the inputs to P_3 and P_4, and the screen voltage of P_4 was adjusted to give zero difference signal; this made the gain of P_4 equal to that of P_3. Input 1 was then removed, and the sum output was made equal to the difference output by varying the screen voltage of P_2. Both inputs were restored and the battery transmitter moved to the E–W line; the only remaining adjustment was for the N and S loop aerials to be adjusted for equality of phase and amplitude.

(5.2) Field Performance

The relative accuracy of the system in respect of both azimuth and elevation depends upon the direction of arrival of the signal. It is determined by the magnitude of the phase difference, $\Delta\phi$ (measured on either aerial pair), corresponding to one degree change in azimuth or elevation of the direction of arrival. The $\Delta\phi$ values will be different for each aerial pair; the average has been used in deriving the results given below and in Figs. 7 and 8. There are marked departures near the N, E, S and W directions, so that the $\Delta\phi$ values are not valid for these azimuths.

For elevation the relative accuracy is maximal $[\Delta\phi(\delta) \simeq 4°]$

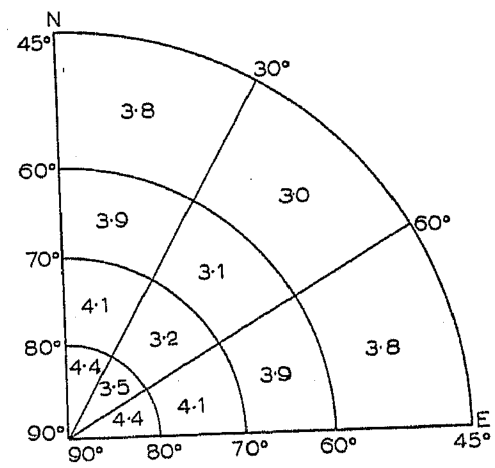


Fig. 7.—Phase differences corresponding to 1° change in elevation, for elevations down to 45° in the N–E octant.

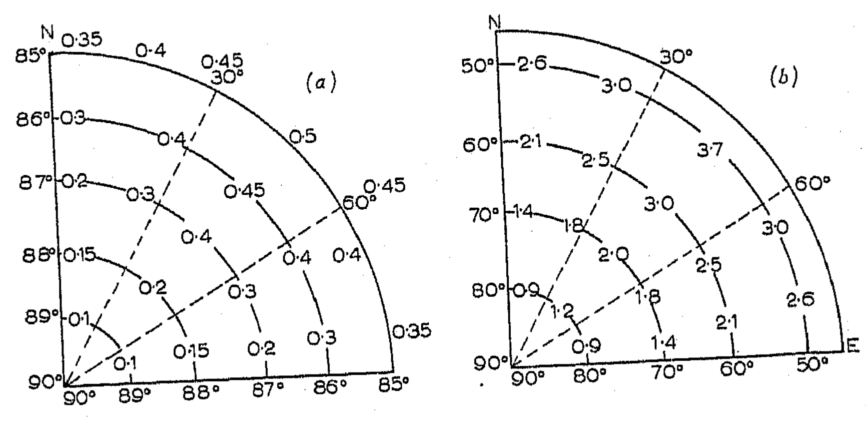


Fig. 8.—Phase difference corresponding to 1° change in azimuth for various elevations in the N–E octant.
(*a*) Elevation angles between 90° and 85°.
(*b*) Elevation angles between 90° and 45°.

for values of δ near 90°, and falls slowly as δ decreases, as shown in Fig. 7.

The position is reversed for azimuth, the phase difference $\Delta\phi(\alpha)$ being zero when $\delta = 90°$. It increases very rapidly out to $\delta = 85°$, and more slowly thereafter. At $\delta = 45°$ the value of $\Delta\phi(\alpha)$ is approximately 3°. These changes are shown in Figs. 8(*a*) and 8(*b*).

The absolute accuracy has been checked with the aid of an aircraft carrying a suitable transmitter. Bearings and elevations of the aircraft flying at 5000 ft were found by direct sighting, using a large Perspex table, and compared with the values found by the radio method. No significant systematic error was apparent in any sector of the observed hemisphere. In particular, no unbalance effect due to power cables being led in along the N–S line was detected. The bearings and elevations were found to be correct to $\pm 2°$ down to the minimum observed elevation of 35°. (This does not preclude the possibility of distant site errors, but it is extremely unlikely that these will be more than one or two degrees for elevation angles greater than 35°.) Since the system was required to record only slow and relatively large changes in direction of arrival, measurement to a higher accuracy was not necessary.

(6) TYPICAL MEASUREMENTS ON REFLECTED PULSES

Two typical records are shown in Figs. 9 and 10, the total range in both being 500 km. The first trace at the left-hand side

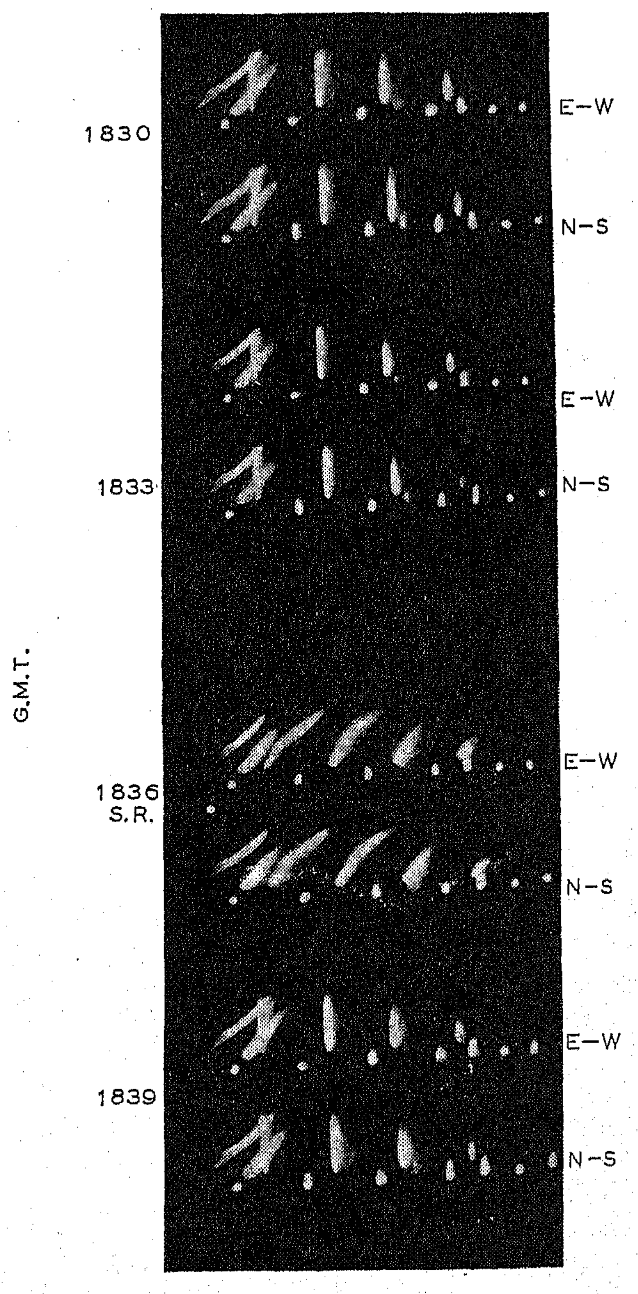


Fig. 9.—Record for the 19th March, 1955, showing zenithal reflections from the E-region.

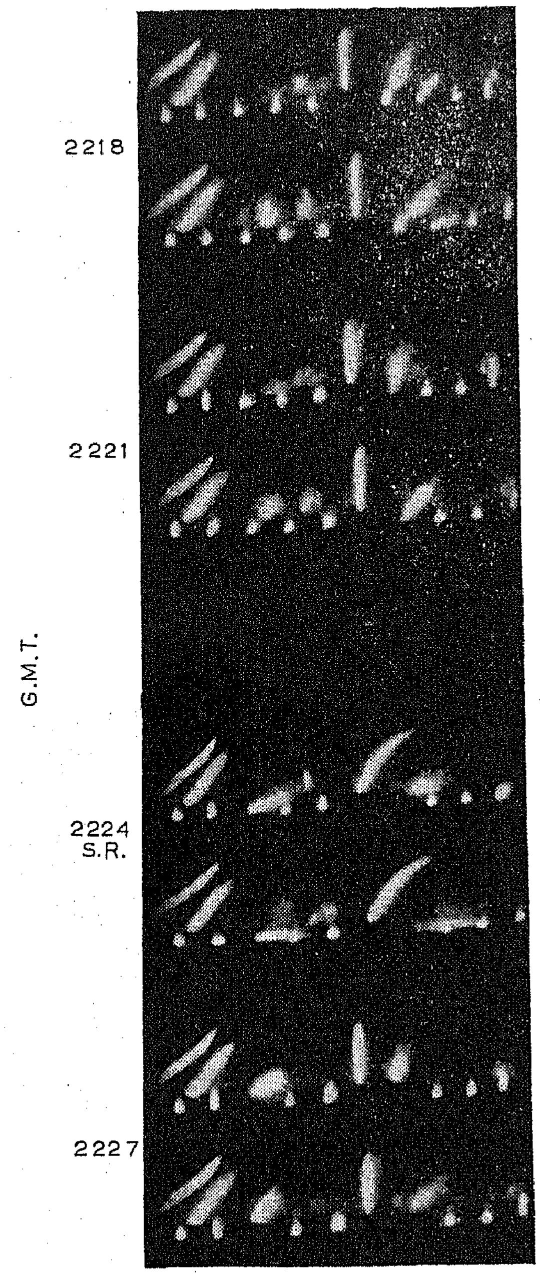


Fig. 10.—Record for the 20th March, 1955, showing non-zenithal reflections from E- and F-regions.

is the transmitter ground pulse, followed by the 50 km range dot, then the signal-generator pulse at 45°. All other traces are either range dots or ionospheric echoes. Fig. 9 is by far the most common type of record and is typical of steady conditions with all echoes coming from the zenith. For zenithal reflections ϕ_1 and ϕ_2 are both zero, so that the echo traces for both aerial pairs are perpendicular to the time-base sweep. During the sense run, with approximately 90° of phase advance inserted in the N and E signals, the recorded traces lie over at an angle of approximately 45°.

A different state of affairs is depicted in Fig. 10. Of the traces shown, those of particular interest are the extra F trace at about 350 km and the Es trace at about 150 km. Accurate range measurements are taken from the associated range record (not shown) and the angle measurements on the phase records are taken with these range measurements as centres. (A finite signal amplitude is necessary before the trace will brighten; this means that there is a slight gap between the base line and the beginning of each echo trace.) An analysis is given below of the two echoes referred to above:

F echo: 2218 hours. Normal run: $R' = 355$ km.*
From E–W trace $|\phi_2| = 50° (= 2 \times 25°)$.
From N–S trace $|\phi_1| = 92° (= 2 \times 46°)$.

* The term virtual range ($R' = P'/2$) is used in preference to the virtual height, h', since the ranges concerned are often slant ranges.

2221 hours, normal run: $R' = 350\,\text{km}$, $|\phi_2| = 46°$, $|\phi_1| = 80°$.
2224 hours, sense run: $R' = 345\,\text{km}$, $|\phi_2'| = 134°$, $|\phi_1'| = 172°$.
2227 hours, normal run: $R' = 340\,\text{km}$, $|\phi_2| = 36°$, $|\phi_1| = 82°$.

From the sense run the signs of all the phase differences may be allocated as positive, e.g. if ϕ_1 were negative the value of $|\phi_1'|$ would have been about 10°. These results and the derived directions of arrival are set out in Table 1 below:

Table 1

F-ECHO DIRECTIONS

Time	R'	ϕ_1	ϕ_2	α	δ
hours	km	deg	deg	deg	deg
2218	355	+92	+50	49	65
2221	350	+80	+46	50	68
2227	340	+82	+36	44	69

Es echo:
2224 hours, sense run: $R' = 150\,\text{km}$, $|\phi_2'| \simeq 154°$, $|\phi_1'| \simeq 176°$.
2227 hours, normal run: $R' = 145\,\text{km}$, $|\phi_2| \simeq 105°$, $|\phi_1| \simeq 70°$.

Thus we have
2227 hours, $R' = 145\,\text{km}$,

$$\phi_1 = +70°,\ \phi_2 = +105°,\ \alpha = 77°,\ \delta = 60°.$$

The accuracy of angle measurement for these particular Es echoes is obviously less than that for the F echoes, but reference to Figs. 7 and 8 indicates that the overall accuracy in direction of arrival is still of the order of $\pm 2°$.

(7) ACKNOWLEDGMENTS

The authors wish to thank Professor H. C. Webster, who made some of the original suggestions for the method of display and maintained a very active interest in the project. Assistance has also been received from Mr. H. T. McGregor and Mr. J. R. Hanscomb. The Royal Australian Air Force co-operated in carrying out the aircraft test of the system.

The work described forms part of the programme of the Radio Research Board of C.S.I.R.O., and is published by permission of the Board.

(8) REFERENCES

(1) ROSS, W.; BRAMLEY, E. N., and ASHWELL, G. E.: "A Phase-Comparison Method of Measuring the Direction of Arrival of Ionospheric Radio Waves," *Proceedings I.E.E.*, Paper No. 1134, July, 1951 (**98**, Part III, p. 294).

(2) ECKERSLEY, T. L., and FARMER, F. T.: "Short Period Fluctuations in the Characteristics of Wireless Echoes from the Ionosphere," *Proceedings of the Royal Society*, A, 1945, **184**, p. 196.

(3) BARFIELD, R. H., and ROSS, W.: "The Measurement of the Lateral Deviation of Radio Waves by Means of a Spaced-Loop Direction-Finder," *Journal I.E.E.*, 1938, **83**, p. 98.

(4) ROSS, W.: "The Development and Study of a Practical Spaced-Loop Radio Direction-Finder for High Frequencies," *ibid.*, 1947, **94**, Part III, p. 99.

(5) ROSS, W.: "Site and Path Errors in Short-Wave Direction-Finding," *ibid.*, p. 108.

(6) HOPKINS, H. G., and HORNER, F.: "Direction-Finding Site Errors at Very High Frequencies," *Proceedings I.E.E.*, Paper No. 773, January, 1949 (**96**, Part III, p. 321).

(7) PIERCE, J. A., MCKENZIE, A. A., and WOODWARD, R. H.: "Loran" (M.I.T. Radiation Laboratory Series, No. 4), p. 367.

(8) WILLIAMS, F. C., and MOODY, N. F.: "Ranging Circuits, Linear Time-Base Generators and Associated Circuits," *Journal I.E.E.*, 1946, **93**, Part IIIA, p. 1188.

(9) DE WALDEN, S., ROCKE, A. F. L., BARRETT, J. O. G., and PITTS, W. J.: "The Development of a High-Frequency Cathode-Ray Direction-Finder for Naval Use," *ibid.*, 1947, **94**, Part IIIA, p. 823.

(10) BRAMLEY, E. N.: "Diversity Effects in Spaced-Aerial Reception of Ionospheric Waves," *Proceedings I.E.E.*, Paper No. 1062, January, 1951 (**98**, Part III, p. 19).

(9) APPENDIX

(9.1) Principles of the Sum-and-Difference Phase Display

In Fig. 11 let A_1 and A_2 represent the two r.f. voltages with a phase difference of ϕ. The sum and difference of the two

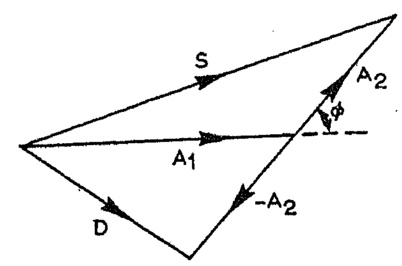


Fig. 11.—Sum-and-difference vector diagram.

vectors are shown in magnitude and direction, and their magnitude is given by

$$S^2 = A_1^2 + A_2^2 + 2A_1A_2\cos\phi$$
$$D^2 = A_1^2 + A_2^2 - 2A_1A_2\cos\phi$$

Thus
$$\frac{|D|}{|S|} = \left(\frac{1 - \dfrac{2A_1A_2}{A_1^2 + A_2^2}\cos\phi}{1 + \dfrac{2A_1A_2}{A_1^2 + A_2^2}\cos\phi}\right)^{1/2}$$
$$= \left(\frac{1 - \cos\theta}{1 + \cos\theta}\right)^{1/2}$$
$$= \tan\tfrac{1}{2}\theta$$

where
$$\cos\theta = \frac{2A_1A_2}{A_1^2 + A_2^2}\cos\phi \quad . \quad . \quad . \quad . \quad . \quad (1)$$

When
$$A_1 = A_2,\ \cos\theta = \cos\phi$$

and
$$\frac{|D|}{|S|} = \tan\tfrac{1}{2}\theta$$

and the phase display will then show a straight line at an angle $\frac{1}{2}\phi$ to the direction of the sum displacement, i.e. to the vertical.

Under simple conditions of no fading A_1 and A_2 will be equal and the tilt of the phase display line will be $\frac{1}{2}\phi$.

(9.2) Diversity Effects

In practice there always is a certain amount of fading in both the amplitude and phase characteristics of the received signals. Both of these effects lead to flicker in the angular position of the phase-display line. The effect of amplitude difference can be seen from eqn. (1). When $A_1 \neq A_2$ then $|\cos\theta| < |\cos\phi|$ and the line will be rotated towards the 45° position (independent of the actual value of ϕ). The effect of variations in phase difference will obviously be to introduce corresponding variations in the angular position of the recorded line (independent of the relative signal amplitudes). Both of these effects will normally be superimposed, and the resultant angular position of the line is thus dependent on both types of fading.

Bramley[10] has derived expressions for the instantaneous amplitude and phase differences in spaced aerials for a number of relevant ionospheric conditions. Using his nomenclature, in each case the distribution of the amplitude difference, $x = A_1 - A_2$, is symmetrical about zero (average amplitudes are equal) and ϕ is symmetrical about a particular value which depends on the direction of arrival of the radio waves.

With the spacings used in this equipment the amplitude correlation coefficient has been found from fading records taken at Brisbane to be between 0·6 and 0·7. Using the lower value as that giving the most severe condition, we obtain the following values for the mean deviations of amplitude and phase:

For a random distribution of rays

$$\overline{|x|} \simeq \frac{2\bar{A}}{\pi}(1-\rho^2)^{1/2} \simeq 0\cdot 5\bar{A}$$

$$\overline{|\phi|} = \arccos\rho \simeq 50^\circ$$

The recorded line will show a sharp vertical edge shading away towards the 45° position, with spreads due to amplitude differences having a mean deviation (in line position) of 11° [obtained by substitution in eqn. (1)], and owing to phase differences, having a mean deviation of 25°. This is obviously an extreme case only met with occasionally in practice.

In the presence of a steady signal with a specular to non-specular (signal/noise) ratio of b

$$\overline{|x|} = \frac{\bar{A}}{b}\left[\frac{2(1-\rho)}{\pi}\right]^{1/2} = \frac{\bar{A}}{b}(0\cdot 5)$$

$$\overline{|\phi|} = \frac{1}{b}\left[\frac{2(1-\rho)}{\pi}\right]^{1/2} = \frac{0\cdot 5}{b}\text{ radian}$$

When $b = 2$—which has been found to hold for a fair proportion of night-time conditions at Brisbane—if the steady signal is from overhead the recorded line will show a sharp vertical edge shading away with amplitude and phase spreads having mean deviations of 6° and 7° respectively. If the steady signal is such as to give a phase difference of 60°, the line will be at 30° to the vertical, with negligible amplitude-difference spread, and phase-difference spread giving a mean deviation of ±7°.

Since about 3° of phase difference is roughly equivalent to 1° difference in direction of arrival (Section 5.2), these spreads of line position will not lead to any uncertainty greater than ±3–4°. This uncertainty is further reduced, since the photographic integration process essentially favours the most probable position of the phase line. Diversity spreading is shown to a limited extent in Figs. 9 and 10.

DISCUSSION ON "AN ATTRACTED-DISC ABSOLUTE VOLTMETER"*

NORTH-WESTERN CENTRE, AT MANCHESTER, 3RD MAY, 1955

Mr. F. W. Taylor: The principles of absolute voltmeters for the measurement of high voltages have long been established, as the References to the paper confirm, and I have often wondered why no one has applied them to produce an accurate instrument which could be employed in any high-voltage laboratory. The accuracy claimed is generally obtained only under certain ideal conditions which do not often occur in the laboratory, and the range, particularly in the upward direction, is distinctly limited.

The instrument described does not fall completely into this category, since the operating voltage is high, it can be made in a good laboratory workshop, and it is transportable and could be used outdoors. The author, however, begins by basing his design on a 500 kV instrument, and ends by checking it only up to 250 kV. Some further work on the design of the throat piece and bushing would increase the operating range up to the design figure of 500 kV. One suggestion would be to increase the diameter of the bushing at the throat and cover the top electrode supporting tube with insulating material of high permittivity where it passes through the throat.

With a longer bushing, however, the temperature error in the gap measurement would be greater, and special care would have to be taken to ensure that the settling of gaskets, the warping or twisting of the main tube (where it is not made of porcelain), and any movement of the bushing top cap did not put the high-voltage electrode too much out of alignment.

The necessity for someone to stand beside the voltmeter in the high-voltage area in order to operate and read the micrometer is a nuisance. Has the author considered any form of remote indication? I presume that in the tests described the control box was mounted close to the tank, but in most high-voltage laboratories the cable run from the high-voltage voltmeter to the control room may be 50 yd or more. The capacitance and insulation resistance of these leads and the associated plugs and sockets always detract from the accuracy of the voltmeter itself, and I presume that they would do so in this case.

Has the author encountered any difficulties in the use of the voltmeter owing to surges caused by brushing or corona on the bushing or parts of the external circuit, and also, is the instrument impervious to damage when a flashover occurs or the circuit is suddenly tripped at maximum voltage?

Mr. E. R. Hartill: In the first part of the paper the author has outlined his reasons for preferring a compressed-gas design to a vacuum enclosure, and this decision has led to a satisfactory design. The difficulties encountered in a vacuum-type instrument, however, are not so much due to leaks as he tends to suggest, since modern techniques are sufficiently well advanced to detect them. It took just over a week to detect all the leaks in an absolute-vacuum voltmeter envelope of about 15 ft^3 volume using a gas probe and a differential Pirani gauge.

The chief difficulty is due to the spasmodic release of gas from electrode surfaces and from the various components, with consequent discharge within the vacuum envelope. This is normally overcome in sealed-off vacuum devices by a preliminary bake-out with subsequent gettering. Such a procedure is inadmissible in the demountable absolute-vacuum voltmeter, where continual adjustment of the internal mechanism is required. In addition, there are difficulties associated with the vacuum bake-out of the delicate mechanism and rather large case, and with the scrupulous cleanliness required. The discharge due to gassing causes flashover of the porcelain electrode supports

* BOWDLER, G. W.: Paper No. 1675 M, July, 1954 (see 102 B, p. 301).

within the vacuum envelope at quite low voltages. However, it should be possible to shield the supports or to grade the voltage distribution along them, so that they operate satisfactorily.

Eqn. (2) is, I believe, based on two-dimensional symmetry, which, of course, does not strictly apply in the case of the circular disc and guard electrodes. If this is so, has the author calculated the magnitude, however small, of the additional error introduced by neglecting this factor?

Mr. A. S. Husbands: An instrument such as that described in the paper should provide a means of improving the general accuracy of high-voltage measurement. The instrument is an absolute one, but it is also a practical one in that it can be used directly as a precision voltmeter in appropriate circumstances. However, the high accuracies of measurement must be accompanied by a corresponding precision in the control and stabilization of the voltages to be measured, and the solution of these problems may involve a greater effort than the provision of the voltmeter itself. In general, it would seem that the absolute electrostatic voltmeter would be most useful in checking and calibrating sub-standard devices, which would be used for the normal measurement work. This procedure would result in a more certain knowledge of the errors involved in the normal measurements—a knowledge which is particularly vague in many of the less-precise present-day measurements. The same calibration principle may be extended indirectly to some instruments which are more versatile than the electrostatic voltmeter (e.g. peak voltmeters).

The precision voltmeter, in conjunction with low-voltage standards, would also enable the determination, at high voltages, of capacitances and resistances, without the need for standard units of high voltage rating.

The main criticism of the instrument described is the suspension of the high-voltage electrode on a long stem. This arrangement must permit quite pronounced vibrations of the electrode, which could be serious in a noisy laboratory or in one subjected to vibration. In addition, the changes of electrode spacing with thermal conditions would require relatively frequent checks of the zero setting of the separation indicator. A solution to this objection would be to support the high-voltage electrode by relatively small insulators mounted inside the tank and in the compressed-gas medium. Then the long bushing could be used as a lead-in only, with a flexible connection to the electrode.

Mr. L. H. A. Carr: The author has put difficulties in the way of those wishing to follow the theory of his instrument by using a multiplicity of unit systems. In the List of Principal Symbols, the first 16 symbols are given in M.K.S. units, and the last two in C.G.S. units. But whereas in this list the symbol of force is stated to be evaluated in newtons, in the body of the paper (e.g. Sections 2.2 and 2.3) this same symbol is used to represent force in kilogrammes.

Mr. W. P. Baker: The effect of vibration of the attracted disc under the pulsating force of a 50 c/s applied field does not appear to have been discussed in the paper. As the author has shown, the force on the disc varies very rapidly as the disc passes through the plane of the guard plate. Under vibrating conditions, therefore, the main force on the disc would be greater for a given spacing and applied r.m.s. voltage than it would be under stationary conditions, and the instrument would tend to read higher on 50 c/s alternating current than on direct current. Has this source of error been considered?

In the majority of applications of high-voltage measurement, e.g. flashover tests and dielectric proof or breakdown tests, the important parameter of the applied voltage is its peak value. There is little advantage in being able to read the r.m.s. value to better than 1% if the crest factor of the voltage wave is more than 1% different from $\sqrt{2}$. More attention should be devoted to the development of better peak voltmeters rather than instruments reading r.m.s. voltage.

Mr. G. W. Bowdler (*in reply*): I am glad to know that Messrs. Taylor and Husbands look upon this as a routine measuring instrument, because that is what it was intended to be rather than a primary standard which tends to be regarded as a museum piece. The adoption of Mr. Taylor's suggestion concerning remote indication of the electrode spacing and of another speaker regarding automatic adjustment of the coil current by means of an amplifier fed from the detector terminals of the audio-frequency bridge would both greatly enhance the utility of the instrument. I see no reason why the length of the leads between the control panel and the instrument should not be as much as 50 yd; the earth capacitance of the leads to L_1 and L_2 (see Fig. 4) would then be about 1 500 $\mu\mu$F, which would not seriously shunt the resistors R_{10} and R_{11}.

The paper is, admittedly, somewhat incomplete without any data about the performance of the instrument with the larger insulator. This has only recently been fitted, and up to the present the maximum 50 c/s voltage which the instrument has withstood has been 350 kV (r.m.s.) when filled with nitrogen at a pressure of 150 lb/in^2. A resistance of about 100 kilohms has been used in the lead to the high-voltage terminal to limit the discharge current when breakdown occurs within the instrument; perhaps because of this no trouble has been encountered with transient over-voltages due to discharges elsewhere in the circuit.

I agree with Messrs. Taylor and Husbands that the method of supporting the high-voltage electrode is one of the least satisfactory features of the design of the instrument; it is the price one has to pay when aiming for 500 kV. The longitudinal stability of the porcelain insulator has nevertheless been very satisfactory, and the lack of lateral rigidity in the mounting of the high-voltage electrode, mentioned in Section 3, has not proved troublesome. A new high-voltage electrode, only one-third of the weight of the original one, has been made for use with the larger insulator, and the conical guide at the upper end of the supporting stem has also been made nearly three times as large as shown in Fig. 1(*a*). In this way, a reasonably rigid mounting of the high-voltage electrode has been obtained with the larger insulator.

I was very interested to hear of Mr. Hartill's experiences with a similar instrument operating *in vacuo*; here again high voltages bring increasing difficulties in their train. Since the charge density on the disc departs from uniformity only over a narrow annulus about 1 mm wide adjoining the edge, the National Bureau of Standards workers, when they were considering the effect on the attractive force of slight departures of the disc from the coplanar condition, were justified in applying the results of an analysis of the case in which the disc is of infinite diameter. I have not estimated the magnitude of any error that this might have on the value of the function f in eqn. (2), but since the instrument is used under conditions in which the term fh in this equation is negligible, it hardly seems necessary.

I must plead guilty to the charges brought forward by Mr. Carr, but unfortunately it was too late to eliminate the inconsistencies he mentioned before the paper was published. The temptation at the end of Section 2.1 to express the attractive force F in terms of the weight of familiar masses was fatal, for beyond this point the forces are repeatedly quoted in kilogrammes or grammes instead of newtons. The gravitational force g in Table 2 and the effective electrode spacing, c, in Tables 2 and 3 should also be expressed in the same units that were assigned to these quantities in the List of Principal Symbols.

Mr. Baker's question has led me to examine theoretically the subject of the vibration of the disc when an alternating voltage is applied to the instrument, and the results are given as an

Appendix to the reply. From data given in the paper, the calculated natural frequency of the disc system is 12·8 c/s. The amplitude of the forced vibration when a 50 c/s voltage stress of 100 kV/cm acts on the disc is 43 microns, and this will cause the mean attractive force to be reduced by 40 and 66 parts in 10^4 at electrode spacings of 5 and 1 cm, respectively; the derived voltages will be in error by half this amount. In the tests recorded in Table 2, the stress acting on the disc ranged from 45 to 64 kV/cm and the voltage errors due to the vibrations of the disc ranged from 4 to 10 parts in 10^4; the a.c. supply during this test was not sufficiently steady for errors of this magnitude to be detected with certainty. The disc vibrations could be reduced either by increasing the mass of the moving system, e.g. using a heavier gauge plate K (see Fig. 3) and thus decreasing its natural frequency, or by opposing the pulsating attractive force with a pulsating balancing force, automatically controlled. The latter would appear to be desirable.

Appendix to Reply

Amplitude and Phase of Forced Oscillations.—Let the disc be driven by a periodic force $F\cos 2\omega t$ due to an applied voltage of frequency* $\omega/2\pi$.

The equation of motion is thus

$$m\ddot{h} + \rho\dot{h} + \mu h = F\cos 2\omega t$$

where m = Mass of disc.
h = Displacement of disc.
ρ = Damping factor.
μ = Stiffness of support.

The solution of which is

$$h = \frac{F}{m}\frac{\cos(2\omega t - \epsilon)}{[(\omega_0^2 - 4\omega^2)^2 + 4\alpha^2\omega^2]^{1/2}} \quad . \quad . \quad . \quad (A)$$

where $\alpha = \rho/m$

$$\omega_0 = \sqrt{(\mu/m)} = 2\pi \times \text{natural frequency of disc}$$

$$\tan\epsilon = \frac{2\alpha\omega}{\omega_0^2 - 4\omega^2}$$

The value of $\omega_0/2\pi$, calculated from the values of μ and m quoted in the paper, is 12·8 c/s. Hence for a 50 c/s applied voltage $\omega \simeq 4\omega_0$, and if the disc is critically damped, $\alpha = 2\omega_0$.

The denominator of eqn. (A) is then $65\omega_0^2$, and the amplitude of vibration is 1/65th of that $(F/m\omega_0^2)$ corresponding to zero frequency of the driving force. Also $\tan\epsilon$ equals $-16/63$, i.e. the motion of the disc lags behind the driving force by nearly half a period.

Effect of Oscillations on Attractive Force.—The instantaneous force F_i acting on the disc, when its displacement above the plane of the guard plate is h and the applied voltage is v, is given by

$$F_i = Kv^2(1 + fh) \quad . \quad \text{[eqn. (2) of the paper]}$$

With $v = V\sin\omega t$

and $h = H\cos 2\omega t$

$$F_i = KV^2\sin^2\omega t(1 + fH\cos 2\omega t)$$
$$= \frac{KV^2}{2}(1 - \cos 2\omega t)(1 + fH\cos 2\omega t)$$

The mean value of the force is $\dfrac{\omega}{2\pi}\displaystyle\int_0^{2\pi/\omega} F_i dt$

$$= \frac{KV^2\omega}{4\pi}\int_0^{2\pi/\omega}[1 + (fH - 1)\cos 2\omega t - fH\cos^2 2\omega t]dt$$
$$= \frac{KV^2\omega}{4\pi}\left[t + \frac{fH - 1}{2\omega}\sin 2\omega t - \frac{fHt}{2} - \frac{fH}{8\omega}\sin 4\omega t\right]_0^{2\pi/\omega}$$
$$= \frac{KV^2}{2}\left(1 - \frac{fH}{2}\right)$$

The fractional error due to the oscillation is therefore $-fH/2$.

The value of H can be calculated from eqn. (1), and values of f are given in Section 2.2 of the paper.

* The periodic force is superimposed on a constant force of the same amplitude (F) which is cancelled during the balancing process.

DISCUSSION ON "THERMIONIC VALVES OF IMPROVED QUALITY FOR GOVERNMENT AND INDUSTRIAL PURPOSES"*

SOUTH MIDLAND RADIO GROUP, AT BIRMINGHAM, 28TH FEBRUARY, 1955

NORTH-EASTERN RADIO AND MEASUREMENTS GROUP, AT NEWCASTLE UPON TYNE, 4TH APRIL, 1955

Mr. J. G. Bartlett (*at Birmingham*): The Government departments responsible for the design of electronic equipment are very grateful to the authors for the extensive work they have done. As the authors mentioned, the initial plan was to reduce to a minimum failures occurring under the shock and vibration conditions found in aircraft and field vehicles. It is reasonable to say that this has been achieved. The equipment designer must also play his part, and one of his problems is to keep bulb temperatures down; here the rating charts mentioned will be of great value. It would also help if the heater power of future valves could be reduced. With compact equipments the temperature of the chassis can rise appreciably and we do feel that if the heater power could be reduced a little it would help.

For static equipment we require long-life valves, and in this connection the work reviewed is of importance. The analyses of failures show that the number of faults in static equipment caused by valves are not greater than those caused by other components. I know of an equipment containing 6 500 valves which has now run for about 2 000 hours, and the valve-failure rates to-day are about 1 in 15 hours, which means one failure per 100 000 valve hours. This equipment uses the conventional type of valve, not the improved ones, and valve faults are no more frequent than those arising from other components.

Finally, I believe that the gas pressure in most valves rises if the valve is run at zero anode current; many valves operate like this for a long time. Do the authors think that the life of the valve is thereby appreciably reduced?

Dr. D. A. Bell (*at Birmingham*): I wonder whether we could

* Rowe, E. G., Welch, P., and Wright, W. W.: Paper No. 1740 R, December, 1954 (see 102 B, p. 343).

be told something of the nature of the experimental nickel-alloy for cathodes. A recent paper* suggested for Class 1 valves an exceptionally pure nickel, comparable, I believe, with experiments carried out by the Post Office using platinum as a base. This eliminates interface trouble, but, on the other hand, the silicon was put there for the definite purpose of reducing the barium oxide. The authors of the paper quoted were content with a very much smaller emission as the price to pay for long valve life; this will offend Mr. Bartlett as one would then need a higher cathode-heating power if the same emission were essential. Can one apply the principle of reducing a lot of barium when the cathode is first activated and hoping this will last the life of the valve? I suspect from what has been said about gas evolution and so forth that one cannot. Secondly, I do not believe that the improvement in the relative drifts in the case of the double triode (Fig. 9) is due to mechanical improvements. Many years ago I was concerned about frequency stability of valve oscillators, and capacitance measurements showed that the changes in input capacitance due to heating were nothing like sufficient to account for the frequency changes. It has since seemed obvious to me that the frequency drift was associated with the change of contact potential, but it appears that the authors are claiming that the whole change in characteristic has been removed by eliminating mechanical drift in the valve. From my own measurements on this type of problem I believe that the drifts cannot be accounted for in terms of mechanical movement of the valve, and the reduction in these irregularities must be due to closer control of qualities of materials and methods of treatment in manufacture of the cathode.

Dr. D. N. Truscott (*at Birmingham*): In the case of the trustworthy valve I notice that the designers have been able to set their own specification:

> "The authors have evolved a practical objective in which the standard required is defined as the highest reliability which can be obtained, commensurate with the ability to manufacture by mass-production methods."

This seems at first sight as though the valve manufacturers are "getting away with it" in writing their own specification, but it is really a recognition that the highest quality is obtained in that way.

I have one particular figure which might be worth quoting in connection with the reliability of standard types of valve. During the 1939–45 War we had a complaint from the Air Ministry that a certain valve was not reliable, since in a 1 000-bomber raid some four bombers returned to base owing to failure of this valve. When we investigated the number of valves of this type installed in each bomber we found it to be 65, and assuming a flight period of four hours, this meant there was one valve failure each 65 000 valve hours. The early failure rate on valves produced by mass production can be incredibly low, and I think that this sometimes leads us to expect an even higher standard of performance from the valve makers in the future, although there may be no good ground for such expectations.

I should like to know whether Fig. 1 is typical or whether it is really a mathematical representation of what the designers would like and which does not necessarily line up with the facts. How do these curves take account of the fact that there are normally catastrophic failures at the beginning of life, and that the failures at the end of life may be due to any one of a number of causes?

Mr. R. C. Harman (*at Birmingham*): Can the authors tell us how to minimize grid emission in valves associated with d.c. circuits?

* EDZMAN, S., and LAGERHOLM, G.: "Modern Long Life Electron Tubes for Telephony Purposes—Some Experience of Life Tests," *Ericsson Review*, 1954, 31, p. 94.

As transmitters are called upon to work on higher and higher frequencies, the valves employed are manufactured with increasingly complex glass-metal envelopes, and the electrode spacing becomes less. Many larger ones are forced-air cooled, which involves a certain amount of buffeting; can the authors say whether the life of such valves is minimized owing to vibration fatigue?

Mr. A. B. Muir (*at Newcastle upon Tyne*): In my opinion, there are a number of instances where the use of electronic circuits would greatly improve the operational flexibility and performance of a power system as a whole. Electronic governors have been developed for hydraulic turbines, and the advantages offered over the conventional fly-ball-operated hydraulic governor cannot be lightly put aside. Electronic voltage regulators have been available for some years, and as their response time is practically zero, they enable the greatest possible response rates to be obtained from excitation systems. However, since their cost is high compared with conventional voltage regulators, they have fallen out of favour. This excessive cost is due to the proven need that one, and sometimes two, complete sets of standby equipment are required with the electronic regulator, owing to the relatively frequent failure of components. The wider adoption of electronic relaying on transmission lines would enable greater quantities of power to be transmitted over systems where the power limit is set by stability considerations rather than by thermal rating.

The types of valve mainly associated with power systems are those grouped in Classes 1 and 4 of Section 3.1. Class 1 valves are used primarily in the carrier and phase-comparison protection equipments, whilst the Class 4 valve finds its greatest application in the primary control circuits of generating equipment. The latter equipment is invariably of a closed-loop servo system with large negative feedback, and as such, the operation of the circuit is largely independent of the valve characteristics. On the other hand the operation of equipment using the Class 1 type of valve is very largely dependent on the ability of the valve to maintain its specified characteristics. I suggest that the probable life of these valves should be determined with these points in mind.

For a Class 1 type of valve the probability of the valve maintaining its characteristics between certain specified limits after a period of say 10 000 hours of continuous operation could be taken as the criterion of reliability. In the case of the Class 4 type of valve, the probability that the valve would remain in an operational condition over the same period could be taken as the criterion of reliability.

It would be of interest to learn from the authors whether any move is being made in the general directions which I have indicated, and whether they consider that it is an economic proposition to produce an histogram for each type of valve.

Mr. D. R. Parsons (*at Newcastle upon Tyne*): In the third lantern slide shown by the authors the selected type of valve had a failure slope which was the same as the ordinary types, except that the latter failed badly during the first 24 hours of use. I should like to know why the authors do not keep all types tested for the extended 24 hours?

Valve manufacturers give a guarantee of valve life which is frequently given, in effect, to a circuit which has not been approved by the circuit-application engineer of the valve manufacturer concerned. Frequently valve manufacturers design the circuits for their own valve applications, but in other cases, in the interest of reliability, I think that they should approve circuits using their products.

I consider that for general communication use 15 000 hours is a satisfactory figure for valve life.

It is frequently stated that the life of indirectly heated valves is extended if their heaters are left on continuously and the h.t. supply switched on only during periods of use. Is this statement true?

May I suggest that designers are reluctant to use valves without bases for direct wiring in the circuit because the valve-holder is so useful as a tag strip for small resistors and capacitors? The integral design of valveholder base may be the answer.

Mr. P. F. Cook (*communicated*): One of the most interesting points of this paper is the method which the authors have used to design highly reliable versions of a device whose history was such that a compromise favouring cheapness at the expense of reliability was the convention. It is significant that they apparently decided at an early stage to direct their efforts to the improvement of the valve from one particular point of view: that of mechanical robustness. It is noteworthy that this has brought in its train desirable improvements in many other characteristics. The paper has significance wider than the immediate concern of electronic valves and thus more information would have been welcome on the decisions which led the authors to select the particular parameter of strength as the one of major relevance.

In the Air Registration Board we have evidence that the improved techniques of valve manufacture have enabled the design of electronic equipment of substantially greater reliability than has hitherto been experienced.

We find that there are three major aspects of ensuring reliability of electronic equipment in civil aircraft. The first is the provision of a valve of suitably high consistency and reliability; the second is that of ensuring that the circuit design is suitable for the valve specified; and the third is that of ensuring that the maintenance work is appropriate in all details.

It is pleasant to record that the first aspect is now being taken care of by the valve manufacturers to an increasingly satisfactory extent. The advance, we feel, has by no means reached the appropriate compromise between cost and the reliability obtained, but the recent improvements are substantial.

Regarding the second point, that of circuit design, it has become increasingly recognized that the promotion of collusion between the circuit designer and valve application engineer *at an early stage in the development of the apparatus* has produced notable improvements. Defect records show that equipment which has been scrutinized in this way has given consistently greater reliability in service. For example, the valve engineer has pressed the equipment designer to exploit the 400 c/s supply so that the heaters can be fed in parallel instead of having to be connected in series-parallel strings to operate from the now-obsolescent 19-volt d.c. supply. We are confident that this kind of co-operation is essential for equipment required to be reliable, and although it does involve extra work there is evidence that this has been mutually beneficial.

The third problem, that of proper maintenance, is probably the most difficult at the present stage. It is very much one of psychology. An individual confronted with the repair of a failed device hopes that the item which has failed is the one which takes the least time to replace. Thus, if there is a plug-in device, be it a valve, sensitive relay or any other component, it is tempting to hope that this is the failed component and to replace it before embarking upon the proper logical tests.

I feel that the problem of whether we have soldered-in valve connections or plug-in valves is mis-stated. Now that the reliability of valves has been improved to a similar order of that of other components, the question should be framed as follows: "Shall valves and components be plug-in devices or shall valves and components be soldered devices?" A statistical analysis of failures which occur in plug-in contacts of all types answers the question affirmatively for the latter. The compromise between cost and reliability to which we are used in domestic radio and television may well mean that in that field the plug-in valve is the optimum solution; but for equipment where economics, and other factors such as airworthiness, dictate reliability, the soldered-in component is unquestionably the best. It will have a considerable effect on maintenance procedure, but I predict that this effect will, in the long run, be in a beneficial direction—not only will it make maintenance organizations insist on the provision of adequate test instruments, but it will also ensure their use.

Messrs. E. G. Rowe, P. Welch and W. W. Wright (*in reply*): The valve industry has been working progressively towards more efficient structures with closer electrode spacings which will fulfil the same requirements as the older types of valve but use less heater power. The size of the valve will depend largely on the power expected from it; thus it is logical to have in any one equipment subminiature, miniature and noval all-glass types for low-, medium- and high-power output respectively. The conditions of operation of the valve will, of course, affect its life, and use at temperatures above those recommended by the manufacturer will result in gassing and grid emission, both of which will tend to shorten the life. Operation at zero anode current in general results in clean-up of gas, but will cause rapid increase in interface resistance. In this country we have carried out extensive tests on the latter problem and can now make valves from special cathode nickels in which this fault is reduced to a negligible amount. Moreover, this improvement has been attained without any loss in efficiency, using established production techniques.

Work done to improve valve performance by attention to gas and grid emission has shown good results in reducing characteristic drift. Better mechanical design of the valve structure and better assembly methods result in less handling, and it is thus possible to make valves which are cleaner and are free of defects which may cause certain parts to distort or overheat during processing. Such approaches on a mass-production basis lead to greater uniformity of the product. This, combined with scientifically conceived exhaust and activation processes, produces finished valves which are more stable electrically—a fact which has been demonstrated.

The life expectancy of valves is governed by the way they are designed, the way they are manufactured and the way they are used. We have shown that the valve manufacturer has done much on the first two items but the co-operation between manufacturer and circuit designer leaves much to be desired. We wish that the latter was more willing to take advantage of the assistance that our applications engineers are only too willing to extend.

621.315.212.029.5 : 621.3.011.21

Paper No. 1891 R
Nov. 1955

THE CHOICE OF IMPEDANCE FOR COAXIAL RADIO-FREQUENCY CABLES

By W. T. BLACKBAND, M.Sc., Associate Member.

(*The paper was first received* 9*th December*, 1954, *and in revised form* 12*th May*, 1955.)

SUMMARY

The choice of characteristic impedance of a coaxial cable is discussed. Among the cable properties considered are attenuation per unit length, voltage rating and power ratings based upon thermal and voltage limitations. A method is given for determining the optimum proportioning to satisfy a specification of two or more cable properties. It is shown that the criterion of cable diameter leads to conclusions little different from those based upon the criteria of dielectric or conductor cross-section. The best choice of impedance is 75 ohms for low-loss air-spaced cables, and 50 ohms for general-purpose thermoplastic cables.

LIST OF SYMBOLS

b = Ratio of wall thickness of the outer to that of the inner conductor.
c = Velocity of electromagnetic waves in free space, 3×10^{10} cm/sec.
D = Internal diameter of the outer conductor, cm.
D_1 = Inner diameter of the protective sheath, cm.
D_2 = Outer diameter of the protective sheath, cm.
D_s = Overall diameter of the complete cable, cm.
d = Overall diameter of the inner conductor, cm.
E = Maximum permissible voltage gradient for the insulant, volts/cm.
f = Frequency, c/s.
G_d = Thermal resistivity of dielectric, thermal ohm-cm.
G_g = Thermal resistivity of ground, thermal ohm-cm.
g = Ratio of effective surface resistivities of outer and inner conductors.
h = Ratio of thickness to diameter for outer conductor.
$H_g(x)$ = Attenuation factor.
H_0 = Thermal dissipation corresponding to a temperature rise of θ° C of the inner conductor, watts/cm.
K = Thermal-dissipation constant for surface in air, watts/cm²/° C.
K_1, K_2 = Form factors for the outer and inner conductors.
l = Depth of cable axis below ground level, cm.
$m = K_1\sqrt{\rho_1}$, $n = K_2\sqrt{\rho_2}$ } Coefficients allowing for the form and resistivity of the outer and inner conductors.
P_v = Power rating based on consideration of voltage, watts.
P_0 = Power rating based on thermal considerations, watts.
S_a = Thermal resistance external to unit length of cable in air, thermal ohms/cm.
S_c = Thermal resistance of unit length of protective sheath, thermal ohms/cm.
S_d = Thermal resistance external to unit length of cable in ground, thermal ohms/cm.
t = Thickness of outer conductor, cm.
V = Rated maximum working voltage, volts (peak.)
$x = D/d$, the ratio of outer to inner conductor diameter.
x_{opt} = Optimum value of x for any given property of the cable.
Z_0 = Characteristic impedance, ohms.
Z_s = Input impedance, ohms.
α = Attenuation per unit length, nepers/cm.
α_c = Attenuation per unit length owing to conductor losses, nepers/cm.
α_{c1} = Attenuation per unit length owing to inner conductor loss, nepers/cm.
α_d = Attenuation per unit length owing to dielectric loss, nepers/cm.
$\tan\delta$ = Power factor of dielectric.
ϵ = Permittivity.
θ_c = Temperature of the inner conductor, °C.
θ_s = Temperature of the outside of the cable, °C.
ρ = Resistivity, ohm-cm.
ρ_1, ρ_2 = Resistivity of outer and inner conductors, respectively, ohm-cm.

(1) INTRODUCTION

The standardization of the characteristic impedance of radio-frequency cables is of importance because it would simplify manufacture and facilitate the design of many items ranging from plugs and sockets to complex line equipment. The purpose of the paper is to study the various factors influencing the choice of characteristic impedance and to present the data on which a wise choice of standard must be based.

The many radio-frequency properties of a cable which must be taken into consideration in determining its characteristics require widely differing characteristic impedances, and for this reason there is no single impedance which is most appropriate to all applications. Thus the phrase "optimum or most suitable impedance" can have no meaning unless further defined, and, in fact, the use of that phrase without qualification merely invites the question "Most suitable for what?" The answer might be "For use at a high voltage," and a further question then arises concerning the criterion by which the suitability is assessed —whether on the basis of cheapness, size or weight. It is quite possible, for instance, that the cheapest design may not necessarily be the smallest or the lightest.

The cable designer is concerned with meeting a given specification of radio-frequency properties in the "best" way. The criterion of size is usually taken in deciding which way is "best," and thus the cable might be designed to have a specified attenuation per unit length with the minimum diameter. While this design would not necessarily give the cable of lowest cost or weight, or that using the least copper or dielectric, it is shown in Section 12.3 that cables designed for minimum diameter approximate closely to those designed with these other criteria in mind. For this reason the discussion in the main part of the paper is based on the cable diameter alone.

In many cases one single radio-frequency property dominates the specification, and in achieving that requirement the others

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
Mr. Blackband is at the Royal Aircraft Establishment.

are met with a comfortable margin. In such cases the design can be based upon the study of this property alone, as has been done in Sections 2–8.

The radio-frequency properties considered in this paper are as follows:

Attenuation per unit length.
Voltage rating.
Power rating with voltage limitation.
Power rating with thermal limitation.
Screening efficiency.
Input impedance of matching stubs of coaxial line.

The most important parameter in the design of the cable is the ratio of conductor diameters, D/d, which is here represented as x. For all the above radio-frequency properties, except screening efficiency, the variation of the property with x is independent of the effective dielectric constant of the cable, and for this reason the analysis is made in terms of x rather than Z_0. The optimum value of x deduced for any particular radio-frequency property leads easily to the corresponding optimum characteristic impedance when the effective dielectric constant of the cable construction under consideration is known.

In cases where the specified values of two or more of the properties of the cable involve comparably stringent limitations on the design, the optimum design can be deduced by the method given in Section 12.4.

It will be seen from Sections 2–8 that the optimum impedances deduced from a study of the radio-frequency properties separately differ widely. In Section 9 the conflicting requirements are assessed and the standards of 50 and 75 ohms advanced for the general-purpose solid dielectric and the low attenuation air-spaced cables respectively.

Except for those of Fig. 6, the graphs presented to illustrate the dependence of cable properties upon the diameter ratio differ from those usually published. Thus Figs. 3 and 4 show the variation in cable diameter with conductor diameter ratio for given cable attenuations or voltage ratings, instead of the variations of these properties for constant cable diameter. The special usefulness of this presentation is shown in Section 12.4 which deals with the problem of designing the minimum-diameter cable when both attenuation and voltage-rating specifications have to be taken into account.

All the graphs are plotted on log/log scales, and therefore the sharpness of the minima on all the graphs is directly comparable.

(2) CONSIDERATION OF ATTENUATION PER UNIT LENGTH

The variation of attenuation with x for a given size of coaxial cable is well known and has been discussed by Green, Leibe and Curtis,[1] Malatesta[2] and Kaden.[3]

It is assumed in the paper that the working frequency is so great that the effective depth of penetration of the currents is small compared with the thickness of the outer conductor and radius of the inner conductor. This condition is fulfilled at frequencies above 1 Mc/s for normal cable constructions.

The attenuation of a cable can be considered in two parts, one being due to dielectric loss and the other to conductor loss. The former is independent of x and will not be considered here, and the latter is given by the equation

$$\alpha_c = \frac{5\cdot 27\ 10^{-7}\sqrt{\varepsilon f}}{\log_\varepsilon x}\left(\frac{m}{D} + \frac{n}{d}\right) \text{ neper per centimetre} \quad . \quad (1)$$

where $x = D/d$, and the coefficients $m = K_1\sqrt{\rho_1}$ and $n = K_2\sqrt{\rho_2}$ allow for the form (K_1, K_2) and resistivity (ρ_1, ρ_2) of the outer and inner conductors respectively (see Section 12.1).

The diameter of a cable having conductor loss α_c is thus

$$D = \frac{5\cdot 27\ 10^{-7}K_2\sqrt{(\varepsilon f \rho_2)}}{\alpha_c}\left(\frac{x+g}{\log_\varepsilon x}\right) \text{ centimetres} \quad . \quad (2)$$

where $g = m/n$.

The term in brackets in eqn. (2) is of general usefulness in the study of optimum impedances. This "attenuation factor" is represented by $H_g(x)$.

D is a minimum for the given attenuation α_c when

$$x(\log_\varepsilon x - 1) = g \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

Thus the value of x_{opt} is dependent upon the form and resistivities of the conductors. Fig. 2 gives the graphical solution of eqn. (3) for values of g from 0 to 6·0.

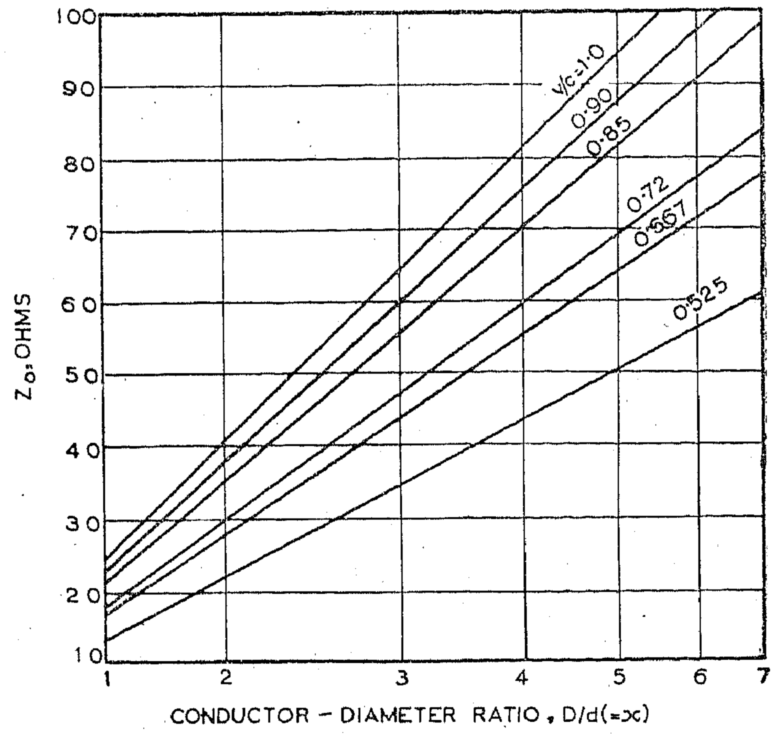


Fig. 1.—Characteristic-impedance variation with conductor-diameter ratio.

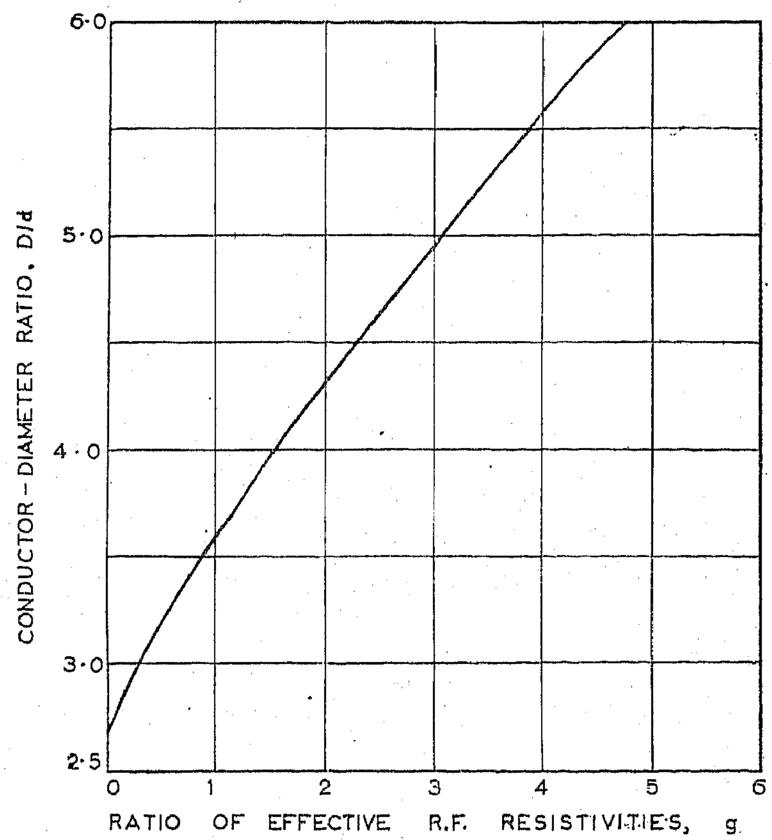


Fig. 2.—Variation of optimum conductor-diameter ratio with the ratio of effective r.f. resistivities for a coaxial cable of given cross-section.

Table 1 shows, for various constructions of outer conductor, the corresponding values of g, on the assumption that the inner conductor is a single copper wire. The Table also shows the values of Z_0 which would be obtained if cables were constructed with these values of x_{opt} with each of various types of dielectric.

Table 1

Optimum Impedances (in Ohms) for Minimum Attenuation

Velocity ratio,* v/c			1·00	0·90	0·72	0·667	0·525
Outer conductor	g	x_{opt}	Z_0 for above v/c values				
Copper tube	1·00	3·59	76	68	55	50	39
Aluminium tube	1·28	3·8	79	71	58	52	41
Lead tube	3·5	5·2	98	88	71	65	51
Plain copper Braid†	2·0	4·3	87	77	63	58	—
	4·0	5·6	103	92	75	68	—

* The values of v/c correspond to the following dielectrics:
1·00 Air.
0·90 Lower limit of so-called air-spaced dielectric.
0·72 Solid polytetrafluorethylene.
0·667 Solid polythene.
0·525 Mineral dielectric.

The values of v/c quoted above are rather higher than would apply to cables completely filled with polythene, polytetrafluorethylene, or mineral dielectric (magnesia). These higher values are representative of cables in current production. The increase in v/c is due to the presence of small air films between the conductors and the body of the dielectric, and air pockets or bubbles within the dielectric. For the mineral dielectric, which has a high permittivity, the effect of the inevitable air pockets around the compacted hard grains of mineral is very pronounced.

† The values of g quoted represent the normal range of relative resistivities of braided screens. The lower limit applies to small cables (0·116 and 0·128 in) and the upper to large cables (0·680 and 0·800 in diameter over dielectric).

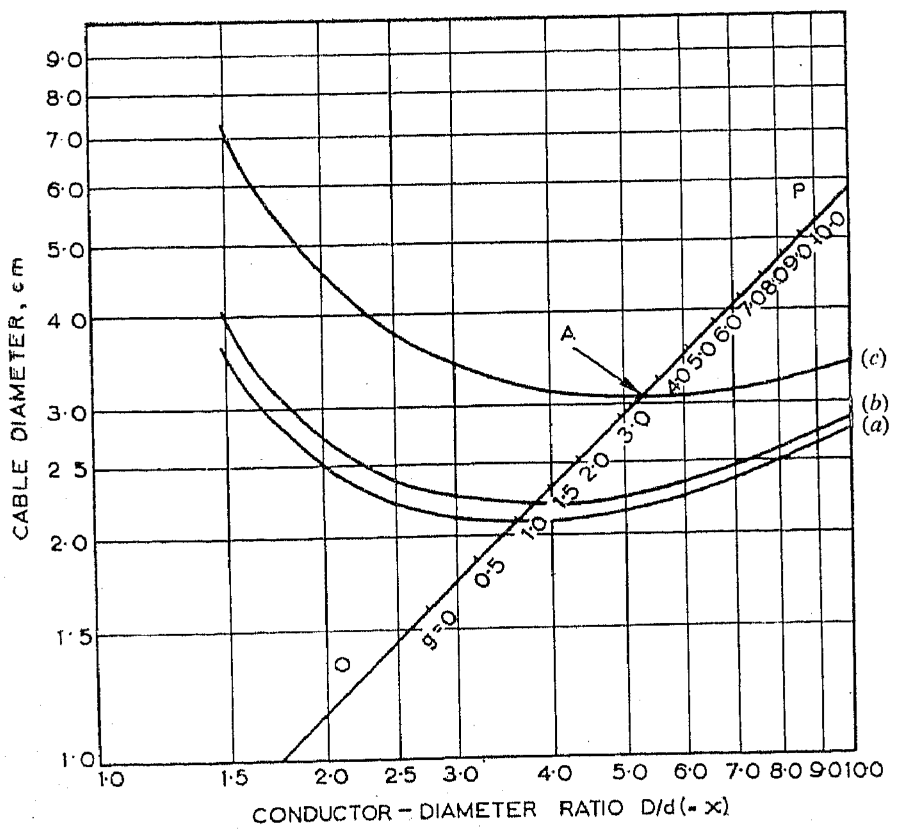


Fig. 3.—Variation of cable diameter with D/d for an attenuation of 1 dB/100 ft at 1 000 Mc/s.

(a) $g = 1\cdot00$.
(b) $g = 1\cdot28$.
(c) $g = 3\cdot5$.

Fig. 3 shows the variation in diameter of the outer conductor with conductor-diameter ratio for constant attenuation ($\alpha_c = 1$ dB/100 ft at 1 000 Mc/s). The lower curve is for a cable with copper inner and outer conductors ($g = 1\cdot0$). Its minimum is at $x_{opt} = 3\cdot59$. Replacing the copper outer conductor by one of aluminium moves the curve up and towards the right, the minimum being at $x_{opt} = 3\cdot8$. The use of a lead outer conductor would move the curve into the position of the top curve, which has its minimum at $x_{opt} = 5\cdot2$. The minima of the curves can be shown to lie on a straight line PQ. The scale marked shows the intersection of the curves with PQ in terms of their values of g.

It will be noticed that the minima of these curves are rather flat, especially for the higher values of g. The effect of dielectric loss, which has not been considered so far, would be to flatten the curves; it cannot move the minima along the x-axis.

(3) CONSIDERATION OF VOLTAGE RATING

For a coaxial cable with homogeneous insulant the rated maximum peak working voltage will be that for which the voltage gradient at the inner conductor reaches the maximum permissible for the particular insulant used.

On the assumption that both conductors are perfect cylinders, this voltage is given by the equation

$$V = \tfrac{1}{2}Ed \log_\varepsilon x \quad \text{volts (peak)} \quad . \quad . \quad . \quad . \quad (4)$$

where E is the maximum permissible voltage gradient for the insulant, and is assumed to be independent of the inner-conductor diameter.* Hence the diameter and rated voltage are related by the equation

$$D = \frac{2V}{E}\frac{x}{\log_\varepsilon x}\,\text{cm} \quad . \quad . \quad . \quad . \quad . \quad (5)$$

Fig. 4 shows the variation in the diameter of the cable with x for $V/E = 0\cdot5$. This would correspond, in an air-spaced line

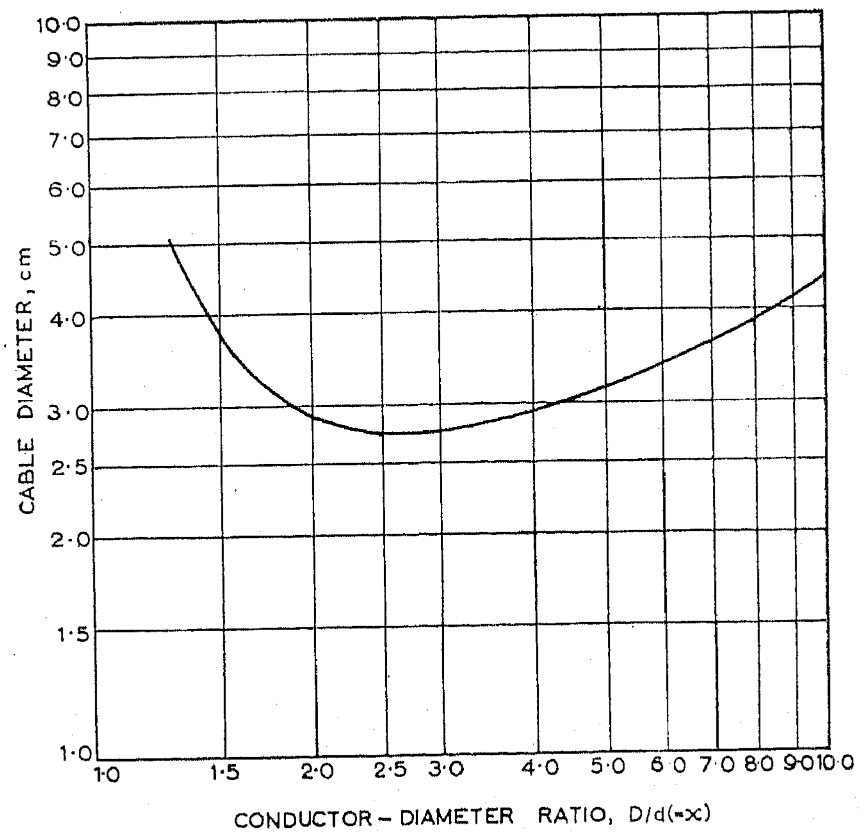


Fig. 4.—Variation of cable diameter with D/d for $V/E = 0\cdot5$ cm.

for which E can be taken as 20 kV (peak)/cm, to a voltage rating of 10 kV (peak). It can be shown by differentiation of eqn. (5) that the minimum diameter will be given when $\log_\varepsilon x = 1$, i.e. for $x_{opt} = 2\cdot72$.

* For air-spaced cables this is not so, the permissible gradient increasing with decrease in inner-conductor size. On the assumption that Peek's law applies, and that the permissible voltage gradient is multiplied by a factor of $\left(1 + \frac{0\cdot301}{r}\right)$ for a conductor of radius r, the values of x_{opt} for the two smallest R.E.T.M.A. standard line sizes ($D = 0\cdot285$ and $0\cdot785$ in) are 3·70 and 3·38 corresponding to $Z_0 = 78$ and 73 ohms, respectively. For larger sizes the values of x_{opt} approach closer to 2·72.

'he characteristic impedance of cables for this value of x_{opt} the types of dielectric recorded in Table 1 are given in Table 2.

Table 2

OPTIMUM IMPEDANCES FOR MAXIMUM VOLTAGE RATING

elocity ratio, v/c	1·00	0·90	0·72	0·667	0·525
ptimum impedance (x_{opt} = 2·72), ohms	60	54	43	40	31

CONSIDERATION OF POWER RATING WITH VOLTAGE LIMITATION

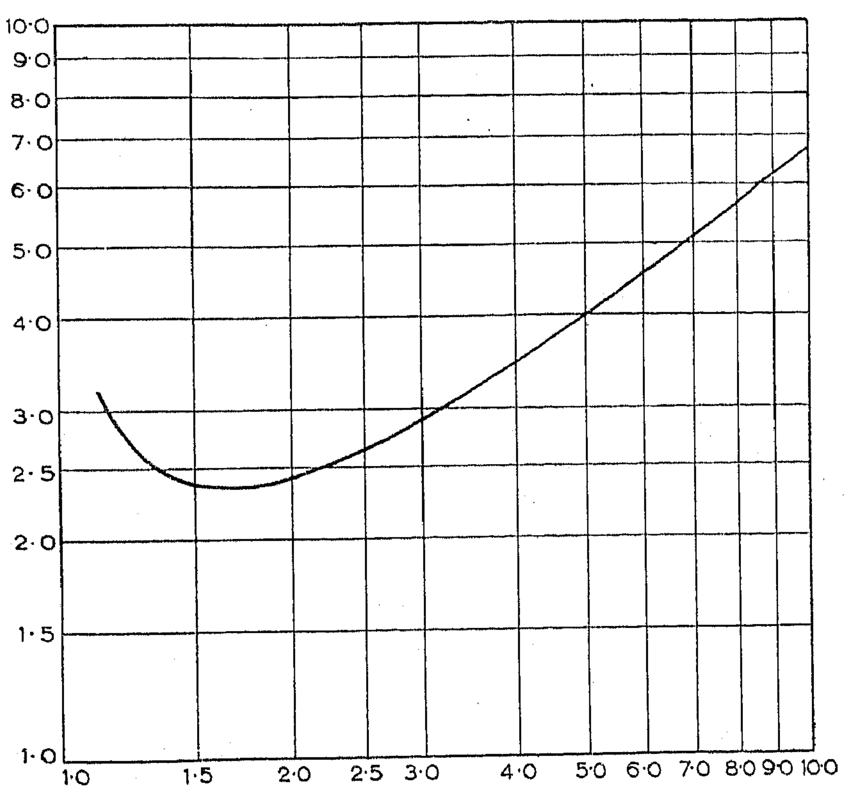
[he peak voltage on a transmission line may be the factor iting the power-handling capacity. It will be least for a given wer level when the line feeds a load equal to its characteristic pedance. When the peak voltage is V (the maximum per- ssible) the power rating is given by

$$P_v = V^2/2Z_0 \text{ watts} \quad . \quad . \quad . \quad . \quad . \quad (6)$$

ls $Z_0 = \frac{60}{\sqrt{\epsilon}} \log_\varepsilon x$ eqn. (5) leads to

$$D = 21\cdot 9\ (\epsilon)^{-1/4}\frac{\sqrt{P_v}}{E}\,\frac{x}{\sqrt{\log_\varepsilon x}} \text{ cm} \quad . \quad . \quad . \quad (7)$$

;. 5 shows the variation of the diameter plotted against x en $\epsilon = 1\cdot 0$ and $\sqrt{P_v}/E = 0\cdot 75$. As this is a multiplying



ʒ. 5.—Variation of cable diameter with D/d for $\sqrt{P_v}/E = 0\cdot 75$ cm and $\epsilon = 1\cdot 0$.

rm its absolute values does not affect the shape of the curve. / differentiation of eqn. (7) it can be shown that the minimum ameter will be given when $\log_\varepsilon x = \frac{1}{2}$ or $x_{opt} = 1\cdot 65$. It ill be noticed that the curve of Fig. 5 rises steeply on either le of the minimum.

The characteristic impedance of cables for this value of x_{opt}, r the types of dielectric recorded in Table 1, are given in ıble 3.

Table 3

OPTIMUM IMPEDANCE FOR MAXIMUM POWER RATING (VOLTAGE LIMITATION)

Velocity ratio v/c	1·00	0·90	0·72	0·667	0·525
Optimum impedance (x_{opt} = 1·65), ohms	30	27	22	20	16

(5) CONSIDERATION OF POWER RATING WITH THERMAL LIMITATIONS

The limit on the thermal dissipation within the cable can arise because of hazards outside or inside the cable. In the former case the temperature of the outer surface of the cable will be lowest when its conductors are proportioned for minimum attenuation, as discussed in Section 2. Unfortunately this case, and the simple answer, are of little more than academic interest. In almost every practical case the limit on thermal dissipation is set by the deterioration of the dielectric where it is in contact with the inner conductor. The power rating is then based upon the maximum inner-conductor temperature permissible for any given dielectric.

The variation of power rating based upon the limitation of the inner-conductor temperature depends upon the construction of the cable, its installation, whether buried or in air, and its working frequency.

The main factors in the cable construction are as follows:

(*a*) The ratio of the effective conductivities of the inner and outer conductors, g.
(*b*) The ratio of the dielectric and conductor losses in the cable.
(*c*) The ratio of the thermal resistivities within and outside the outer conductor.

Of these the first is the most important. Fig. 7 shows a set of curves for solid polythene cables of 0·285 in diameter over dielectric and 0·400 in overall diameter buried in the ground at a depth of 18 in. It will be seen that x_{opt} increases somewhat with increase in g from 1·0 to 4·0, varying from about 2·3 to 3·1 with a consequent variation in Z_0 of 33 to 46 ohms. The ratio of dielectric and conductor losses has only a small effect. Curves (*b*), (*d*) and (*e*) of Fig. 6 illustrate this. Curve (*b*) is for a normal cable at 1000 Mc/s, curve (*d*) shows how little effect there would be on x_{opt} if the dielectric loss were zero, and curve (*e*) shows the effect of increasing the power factor to five times the normal maximum. Because the ratio has little effect the data calculated for a power factor of 0·004 at a frequency of 1000 Mc/s can be applied to other solid polythene cables without introducing appreciable error.

The ratio of thermal resistivities within and outside the outer conductor depends upon the presence of thermally resistant protective layers covering the conductors, and whether the cable is buried in the ground or installed in air. The effect of a p.v.c. sheath upon the value of x_{opt} is small. For the $g = 1$ curve of Fig. 7 it would change x_{opt} from $2\cdot 2_7$ to $2\cdot 4_2$ (Z_0 from 33 to 35 ohms). The curves calculated for cables in the ground and in air are very similar. Curves (*c*) and (*f*) of Fig. 6 relate to the same cable under these two conditions. The difference in x_{opt} is again small. For these reasons the curves of Fig. 7, which are calculated for unsheathed cables installed in the ground, can be taken without serious error as representative also of solid polythene cables having p.v.c. sheaths and installed in air. The figures for power rating (thermal) quoted in Table 4 are advanced on the above basis.

A more detailed consideration of these factors is given in Section 12.2.

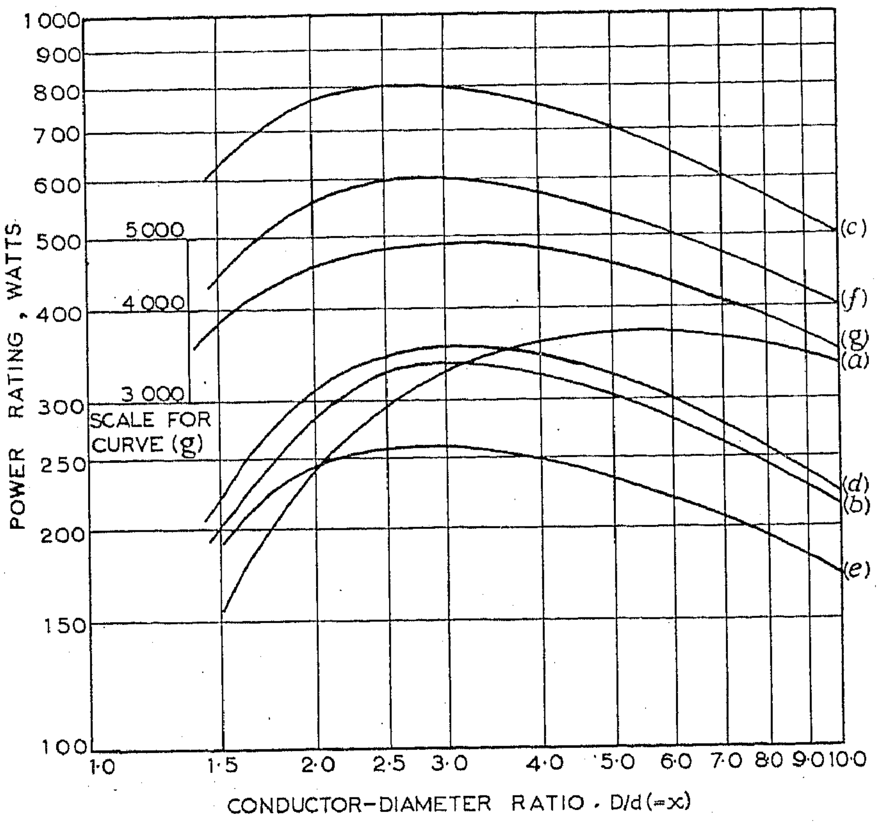


Fig. 6.—Variation of power rating (thermal limitations) with D/d for a given cable diameter.

(*a*) Cable for which $S_g > S_d$.

Lead-sheathed cable of 0·285 in diameter over dielectric installed in ground.

(*b*) Power factor 0·0004.
(*d*) Power factor 0·0000.
(*e*) Power factor 0·0020.

Lead-sheathed cable of 0·800 in diameter over dielectric.

(*c*) Installed in ground.
(*f*) Installed in air $\theta_c = 85°$C.
(*g*) Installed in air $\theta_c = 200°$C.

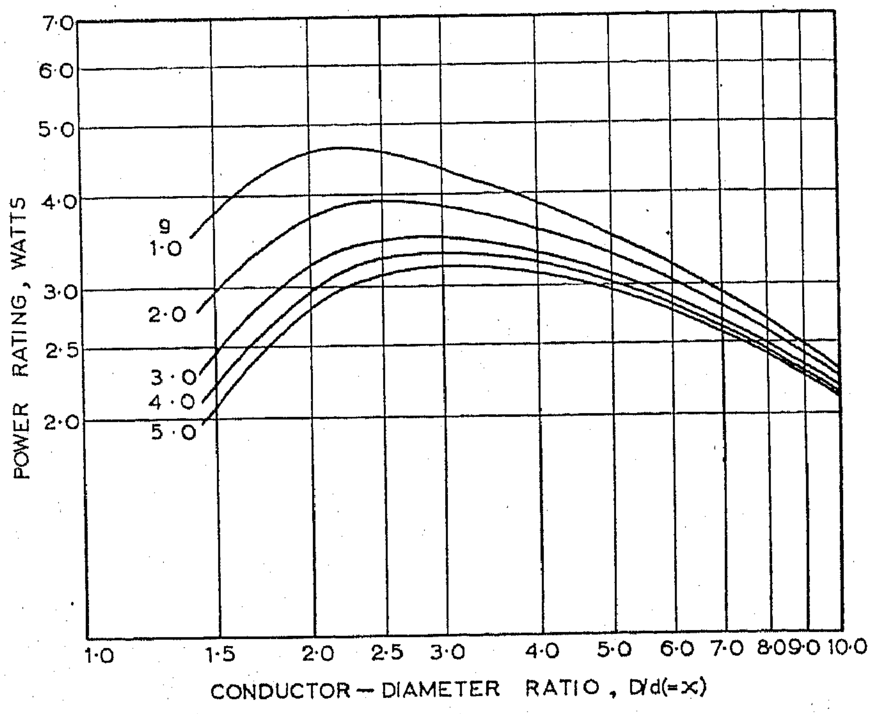


Fig. 7.—Effect of g upon the curve of power-rating/(D/d).

$g = 1·0$ Copper-tube outer conductor.
$g = 2·0$ Braid outer conductor of best construction.
$g = 3·0$ Average braid outer conductor of average construction.
$g = 3·5$ Lead outer conductor.
$g = 4·0$ Braid outer conductor of poor construction.

(6) CONSIDERATION OF SCREENING EFFICIENCY

The radio-frequency screening efficiency expressed in terms of external field for given power transfer is a function of Z_0 (and so of x), but it does not have an optimum value. As the external field is proportional to the current flowing in the outer conductor, it follows that for any one outer conductor transmitting a given power the field will be reduced as Z_0 is increased. However, as the range of Z_0 which can be easily realized for coaxial cables is limited to about 4 : 1, and the variation between the efficiencies of various constructions of screens is much greater than this, it is not likely that considerations of screening would, in practice, influence the choice of diameter ratio of a cable.

(7) CONSIDERATION OF STUB IMPEDANCE

In some applications cables are required to act as high-impedance stubs. If a short-circuited quarter wavelength of cable is to have an input impedance of Z_s, it can be shown that the necessary diameter is

$$D = Z_s B \epsilon \sqrt{\frac{\rho}{f}} \frac{H_g(x)}{\log_e x}$$

where B is a constant, and ρ is the resistivity of the conductors. For a cable of $g = 1$ the curve of D against x has a very flat minimum with $x_{opt} = 9·2$. For greater values of g the values of x_{opt} are even greater.

For a low-impedance stub the input impedance of an open-circuited quarter wavelength of cable will decrease indefinitely as x decreases, and so there will be no optimum value for x.

Where a stub is switched to present either a high or low impedance, the ratio of the two impedances is greatest when the attenuation is least (highest Q-factor), and the optimum ratios are those considered in Section 2.

It will be noted that the consideration of stub impedances does not point to any single preferred value of x. One type of stub needs x to be great, one needs x to be as small as possible, and a third requires the values of x which would be aimed at on consideration of attenuation alone.

(8) UNIFORMITY OF CHARACTERISTIC IMPEDANCE

One of the most important properties of cables for use at centimetric wavelengths is the uniformity of characteristic impedance. This is because local irregularities in Z_0 cause reflections of power which may be serious where long lengths of cable are used. These irregularities in Z_0 can arise in three ways.

(*a*) Variations in diameter of inner conductor.
(*b*) Variations in diameter of outer conductor.
(*c*) Variations in effective permittivity.

In general, variations of the inner-conductor diameter are small, as wire drawing is an accurate process, at least so far as consistency within a short length is concerned. The second variation is much more important in practice. It arises with polythene-dielectric cables because of variations in the diameter of the extruded core. If the core can be made to a given dimensional tolerance the effects of these variations on the characteristic impedance of the cable will be less the greater is Z_0. For a 75-ohm cable the percentage variations will be two-thirds of that for a 50-ohm cable.

The third variation, that of effective permittivity, also has less effect at greater than lower values of Z_0. This is because the chief cause of this variation is the air film which forms round the inner conductor, or odd air bubbles trapped in the dielectric. The effect of these defects will be less in the thicker dielectrics of the higher-characteristic-impedance cables than in the thinner dielectrics of cables with lower characteristic impedance.

For these reasons British practice has been to favour 75 ohms characteristic impedance for solid polythene cables to be used at centimetric wavelengths.

(9) THE CHOICE OF A STANDARD IMPEDANCE

It will be seen from the previous Sections that there is no simple optimum value of Z_0. The cable properties are optimized

Table 4

COMPARISON OF OPTIMUM CHARACTERISTIC IMPEDANCE (IN OHMS)

Property considered in choosing the optimum	Air-spaced cables (copper inner)			Polythene cables $v/c = 0\cdot667$ (copper inner conductors)					
	Pb (outer)	Cu (outer	Al (outer)	Pb (outer)	Al (outer)	Cu (outer)	Plain copper wire braid outer conductor		
	v/c 1·00	$v/c = 1\cdot00$	$v/c = 0\cdot90$				$g = 2\cdot0$	$g = 3\cdot5$	$g = 4\cdot0$
Attenuation	98	76	71	65	52	50	58	65	68
Power rating (thermal)				44		33	37	44	46
Voltage rating	60	60	53	39	39	39	39	39	39
Power rating (voltage)	30	30	27	20	20	20	20	20	20

for distinctly different diameter ratios, and even for a given property the optimum value may vary greatly with differences in construction. The variation of optimum Z_0 for various constructions of cable is shown in Table 4.

Of the classes of cable referred to in Table 4, the most important are as follows:

(*a*) *The completely air-spaced.*—Some cables approach this class having v/c as high as 0·95–0·98.

(*b*) *The partly air-spaced.*—Many transmitter cables have $v/c = 0\cdot90$.

(*c*) *The solid polythene cables with $v/c = 0\cdot667$.*—This class forms the bulk of the flexible and semi-flexible cables.

characteristic impedance. These 75-ohm cables are also favoured for centimetric-band work because of their uniformity of Z_0.

For the general-purpose cables all the radio-frequency properties may be taken as of comparable importance. In Table 5 are shown the percentage excesses of cable diameter over the optimum for each of the four main properties considered here, for $Z_0 = 75$, 60, 50 and 40 ohms. The figures are for solid polythene cables with braided outer conductors having $g = 3\cdot5$. This construction is taken as typical of this class of general-purpose cable. It will be seen from this Table that 50 ohms is the best compromise for this class of cable.*

Table 5

EXCESS OF CABLE DIAMETER ABOVE OPTIMUM FOR VARIOUS STANDARD IMPEDANCES

This Table relates to lead-sheathed or braided solid polythene cables for which $g = 3\cdot5$.

	Percentage excess of cable diameter above optimum			
Cable property	$Z_0 = 75$ ohms	$Z_0 = 60$ ohms	$Z_0 = 50$ ohms	$Z_0 = 40$ ohms
Attenuation	2	1	4	13
Power rating (thermal)	25	8	2	0
Voltage rating	28	10	4	0
Power rating (voltage)	114	63	38	20

Table 4 shows the optimum characteristic impedances for these three classes of cable, based on consideration of various cable properties.

From this Table some deductions can be made about the choice of optimum impedance. The present-day radio-frequency cables can for this purpose be divided into two groups—those for which transmission efficiency is all-important and those for general-purpose use. This classification is helpful because there is also a difference in construction between the two groups. Those which are made to have low attenuation are as nearly air-spaced in construction as possible, and the second group mostly have solid polythene dielectrics with lead or braid outer conductors.

Considering first the low-attenuation cables, it is clear that, while each construction has its own optimum value of Z_0, the only optimum of importance for cables of very low attenuation is that for the air-spaced cables which have values of v/c of nearly unity. For $v/c = 1\cdot0$ and $g = 1$ the optimum value of Z_0 is 76 ohms, but as v/c is always slightly less than unity (0·95–0·98) the optimum value of Z_0 in practice is 75 ohms. Thus if low attenuation is of prime importance, an air-spaced cable should be chosen, and should have $Z_0 = 75$ ohms. Low-loss cables are much used for carrier telephony and the 75-ohm standard is well established in this field. The use of these cables necessitates the production of small flexible and semi-flexible cables for connections between carrier sections and use in associated equipment, all of which must be of 75 ohms

Table 6 shows corresponding percentage excesses for air-spaced cables of $v/c = 0\cdot90$ and aluminium outer conductors ($g = 1\cdot28$). In this case it is doubtful whether 60 ohms offers

Table 6

EXCESS OF CABLE DIAMETER ABOVE OPTIMUM FOR 60- AND 50-OHM IMPEDANCES

The above figures apply to a cable of $v/c = 0\cdot90$.

	Percentage excess of cable diameter above optimum	
Cable property	$Z_0 = 60$ ohms	$Z_0 = 50$ ohms
Attenuation	2	7
Power rating (thermal) ..	8	2
Voltage rating	1	0
Power rating (voltage) ..	27	13

any great advantage over 50 ohms. The 60-ohm cable of this type does have a slight advantage over a 50-ohm cable in attenuation,† but if attenuation is the main consideration this construc-

* For small cables, where flexibility is important, it is usual to strand the inner conductors of 50-ohm cables in circumstances where it would be unnecessary to strand the inner conductor of a 75-ohm cable. This stranding increases the loss in the inner conductor by 20–25% (which corresponds to an increase in total loss of perhaps 7–12% depending upon cable construction and operating frequency). For the larger 50-ohm cables, such as HK50–17, KX50GMI, RG–17/U and Uniradio No. 74, the inner conductors are not stranded, and so this complication does not arise.

† Gutzmann[10] bases his choice upon the criterion of minimum attenuation, and considers the power and voltage-rating criteria to be of much less importance. On this basis he favours the 60-ohm standard.

tion should not be used at all, since the use of a 75-ohm air-spaced cable of $v/c = 0 \cdot 95$ would offer a reduction of attenuation of 5% below that of the $v/c = 0 \cdot 90$, 60-ohm cable of identical size.

Thus the conclusion is reached that for general-purpose use the standard value of Z_0 should be 50 ohms, and that 75 ohms should be a secondary standard for those special applications where low attenuation is of prime importance, or where advantage can be taken of the greater longitudinal uniformity of 75-ohm cables.

(10) ACKNOWLEDGMENT

Acknowledgment is made to the Chief Scientist, the Ministry of Supply, and the Controller of H.M. Stationery Office for permission to publish the paper.

(11) REFERENCES

(1) GREEN, E. I., LEIBE, F. A., and CURTIS, H. E.: "The Proportioning of Shielded Circuits for Minimum R-F Attenuation," *Bell System Technical Journal*, 1936, **15**, p. 248.

(2) MALATESTA, S.: "Condizioni di minima attenuazione mei cavi coassiali per alte frequenze," *Alta Frequenza*, 1940, **9**, p. 755.

(3) KADEN, H.: "Die Dampfung und Laufzeit von Breitbandkabeln," *Archiv fur Electrotechnik*, 1936, **30**, p. 691.

(4) MILDNER, R. C.: "The Power Rating (Thermal) of Radio Frequency Cables," *Journal I.E.E.*, 1946, **93**, Part III, p. 296.

(5) HANSELL, C. W., and CARTER, P. S.: "Frequency Control by Low Power Factor Line Circuits," *Proceedings of the Institute of Radio Engineers*, 1936, **24**, p. 597.

(6) RACOVITCH, O. S.: "The Determination of the Optimum Ratio of Conductor Diameters for a Coaxial Cable," *Elecktrosvyaze*, 1941, **2**, p. 71. A translation into German was printed in *E.T.Z.H.*, 1944, **17/18**, p. 170.

(7) SHINIBEROV, P. YA.: "Coaxial Cable and its Optimum Design," *Priroda*, 1949, No. 1, p. 69.

(8) BLACKBAND, W. T., and BROWN, D. R.: "The Two-Point Method of Measuring Characteristic Impedance and Attenuation of Cables at 3 000 Mc/s," *Journal I.E.E.*, 1946, **93**, Part IIIA, p. 1383.

(9) SMITH, P. H.: "Optimum Ratios for Coaxial Transmission Lines," *Electronics*, February, 1950, p. 111.

(10) GUTZMANN, F.: "Zur Wahl des rectigen Wellenwiderstandes von Hochfrequenz-Kabeln," *Fernmeldetechnische Zeitschrift*, 1954, **7**, p. 3.

(12) APPENDICES

(12.1) The Ratio of Effective Surface Resistivities of the Conductors

The d.c. resistances of two conductors of the same dimensions but different materials will be proportional to their resistivities, ρ_1 and ρ_2. At radio frequencies the current will be concentrated in a surface layer of depth proportional to $\sqrt{\rho}$. At a frequency where this depth is small compared with the conductor dimensions the ratio of the surface resistivity of the two conductors would be proportional to the square root of their resistivities, i.e. $g = \sqrt{(\rho_1/\rho_2)}$.

If the two conductors were of materials of the same resistivity and both of the same external dimensions but of different surface form (e.g. one a cylinder and the other formed of stranded wires or of wire braid) their resistances would differ. Their surface resistivities would be the same, but their effective surface resistivities would differ because of the differences in surface form. The effect of surface form is represented by a "form factor" (K_1, K_2) which relates the effective surface resistivity of a conductor to that of a cylindrical conductor of the same material and the same external dimensions.

Thus the ratio of effective surface resistivities is given by

$$g = \frac{K_1}{K_2}\sqrt{\frac{\rho_1}{\rho_2}}$$

At the broadcasting frequencies the use of a Litz stranding of fine enamelled wires can give a reduction of resistance compared with a solid conductor, because the effective depth of penetration of the current is increased. For such a case K_2 would be less than unity. For some broadcast transmitter cables K_2 at 1 Mc/s is claimed to be as low as $0 \cdot 65$.

In all other cases K_1 and K_2 are greater than unity. For a 7-strand conductor at about 200 Mc/s, a value of $1 \cdot 20$–$1 \cdot 25$ is generally accepted. Kaden[3] has published values of K_1 for outer conductors formed by helically applied tapes. The constructions considered are of fairly long lay and for use at carrier frequencies. Form factors deduced from Kaden's data are $4 \cdot 1$ and $1 \cdot 6$ for tapes making angles of 50° and 70° with the circumference of the cable. At decimetre and centimetre wavelengths with short-lay tapes the form factor may become very great, owing to internal resonances. In one published case[8] a peak of attenuation corresponding to $K_1 = 120$ was measured for a tape at 3 000 M/cs.

Mildner[4] has quoted the following form factors of wire-braided conductors. These were representative of the form factors achieved by Service type approved cablemakers in 1946.

Cable type	UR 17	UR 34	UR 1	UR 4	UR 32
Internal diameter of braid, in	$0 \cdot 800$	$0 \cdot 625$	$0 \cdot 330$	$0 \cdot 285$	$0 \cdot 128$
K_1	$4 \cdot 0$	$3 \cdot 0$	$2 \cdot 25$	$2 \cdot 0$	$1 \cdot 75$

Since 1946 the values of K_1 have been decreased slightly for the larger cables following the introduction of braiding machines capable of carrying more spindles.

So far, attempts to correlate K_1 with the covering factor and lay angle of the braids have led to qualitative rather than quantitative results.

(12.2) Detailed Consideration of Power Rating with Thermal Limitations

The power rating of a cable may be limited by the rise in temperature which can be permitted without causing deterioration of the cable. In most cases the limit is set by the material of the dielectric which may melt, soften or decompose when heated. In such cases this deterioration limits the temperature of the inner conductor, which is the hottest part of the cable.

The temperature of the inner conductor is determined by two factors—the rate at which heat is generated as the r.f. power is attenuated and the thermal resistance of the cable. These two factors are somewhat interdependent because the heat generated is distributed between the inner and outer conductors, and the dielectric, and because, while all the heat traverses the thermal resistance external to the outer conductor, only a part of the heat traverses the thermal resistance of the dielectric.

A detailed account of the methods of assessing the power rating with thermal limitations, P_0, has been given by Mildner.[4] The equations quoted below have been taken from his paper.

It can be shown that, if H_0 is the thermal dissipation of the cable when carrying its rated power,

$$P_0 = H_0/2\alpha \text{ watts} \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

and it is necessary to consider H_0.

The optimum value of x for P_0 depends upon many more

factors than can be considered in the paper, and in order to simplify the study of P_0 we will consider only cables with solid dielectrics. For these cables the following cases are considered:

(a) Lead-sheathed unserved cable buried in ground.
(b) Lead-sheathed unserved cable installed in air.
(c) Braided conductor cable with p.v.c. sheath installed in air.

For a cable buried in the ground

$$H_0 = \theta_c \Big/ \left(S_d \frac{\alpha_{c1} + \frac{1}{2}\alpha_d}{\alpha} + S_c + S_g \right) \quad . \quad . \quad . \quad (9)$$

where θ_c = Rated temperature rise of the inner conductor above ambient.

S_d = Thermal resistance per unit length of the dielectric

$$= \frac{G_d}{2\pi} \log_\varepsilon x \text{ thermal ohms} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (10)$$

S_c = Thermal resistance per unit length of the cable covering

$$= \frac{G_c}{2\pi} \log_\varepsilon \left(\frac{D_2}{D_1} \right) \text{ thermal ohms} \quad . \quad . \quad . \quad . \quad . \quad . \quad (11)$$

S_g = Thermal resistance per unit length of the ground

$$= \frac{G_g}{2\pi} \log_\varepsilon \frac{4l}{D_s} \text{ thermal ohms} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (12)$$

In the above equations, G_d, G_c and G_g are the effective thermal resistivities of the dielectric, the cable covering and the ground respectively; D_1, D_2 and D_s are the inner and outer diameters of the cable covering and the overall diameter of the cable; and l is the depth of the axis of the cable below ground level. The symbols α_{c1} and α_d represent the parts of the attenuation attributable to the inner conductor and dielectric.

(12.2.1) **Lead-sheathed Cables in the Ground.**

With a lead-sheathed cable the thermal resistance of the sheath is relatively small. Thus S_c can be neglected, and the power rating is given by

$$P_0 = \theta_c / 2\alpha \left(S_d \frac{\alpha_{c1} + \frac{1}{2}\alpha_d}{\alpha} + S_g \right) \quad . \quad . \quad . \quad (13)$$

In order to reveal the dependence of P upon x the equation is rewritten as follows:

$$P_\theta = \theta_c \Bigg/ \left[\frac{B}{D} H_g(x) \left(\frac{G_d}{2\pi} \log_\varepsilon x \frac{\frac{x}{g}\frac{\alpha_{c1}}{\alpha_d} + \frac{1}{2}}{\frac{\alpha_c}{\alpha_d} + 1} + \frac{G_g}{2\pi} \log_\varepsilon \frac{4l}{D_s} \right) \right] \quad . \quad . \quad . \quad . \quad . \quad (14)$$

It will be seen that P_θ depends upon x in a complicated way, and that the shape of the P_0/x curve will depend upon several variables. However, the shape of the curve will not depend upon the values of θ_c, B or D, which are multiplying constants. The shape does depend upon $H_g(x)$ and the ratio of the two terms within the curved brackets. This leads to the conclusion that the shape of the curve is dependent upon the following cable parameters:

(a) Conductor form and resistivity, g.
(b) Ratio of thermal resistivities of dielectric and ground, G_d/G_g.
(c) Ratio of attenuations due to the conductors and the dielectric.
(d) Ratio of the depth underground to cable diameter, l/D_s.

The dependence of the curve shape upon these parameters will now be considered.

(12.2.1.1) *Conductor Form and Resistivity.*—In the special case where the resistance S_g external to the cable is large compared with S_d, the term in the curved brackets of eqn. (14) is dominated by its second part [$S_g = (G_g/2\pi) \log_\varepsilon 4l/D_s$)] and can be considered constant with respect to x. In this case the shape of the P_0/x curve is determined by $H_g(x)$ alone. It will, in fact, be that of the curve of Fig. 3 inverted, the variation in conductor material and form introducing the same variations in shape as those previously studied for Fig. 3. Such a curve for a lead-sheathed cable ($g = 3 \cdot 5$) is plotted as curve (a) of Fig. 6. For such curves the maximum power rating for any value of g will be those given for minimum attenuation in Fig. 2.

In general the curved-bracket term is not dominated by a large constant term and so it affects the shape of the curve. The effects of varying the parameters in the curved-bracket term will be considered below. For convenience, the variations will be considered for a medium-size cable (7·25 mm) with lead sheath (such as Uniradio Nos. 45 and 58).

For the calculation of curve (b) of Fig. 6 the following values of the parameters were assumed:

f	g	θ_c	$\tan\delta$	G_d	G_g	D	D_s	l
1 000 Mc/s	3·5	30° C	0·0004	450	120	0·724 cm 0·285 in	1·016 cm 0·400 in	45·7 cm 18 in

The difference in shape between curves (a) and (b) of Fig. 6 shows the effect of the curved-bracket term for this type of cable. The maximum value is displaced to a lower value of x, and the sharpness of the maximum is less marked. The displacement of the maximum is a result of the variation of S_d with x, because for the lower values of x, S_d is less, and a greater heat flow is possible for the same temperature difference across the dielectric.

In Fig. 7 are plotted curves of P_0 against x for this same medium-size cable with a series of outer conductors to give $g = 1, 2, 3, 3 \cdot 5$ and 4. The first corresponds to a solid polythene cable with a copper-tube outer conductor, while $g = 2$ and $g = 4$ represent the limits of form factor with wire-braided outer conductors. The optimum diameter ratios and characteristic impedances are tabulated below:

g	1·0	2·0	3·0	3·5	4·0
x_{opt}	$2 \cdot 2_7$	$2 \cdot 5_3$	$2 \cdot 8_5$	$2 \cdot 9_9$	$3 \cdot 1_3$
Z_0, ohms	33	37	42	44	46

(12.2.1.2) *Ratio of the Thermal Resistivities of the Dielectric and Ground* (G_d/G_g).—The smaller this ratio the less is the variation of the curved-bracket term with change in x. The pair of curves (b) and (c) of Fig. 6 correspond in shape to a pair of cables with $G_d/G_g = 3 \cdot 75$ and $4 \cdot 55$, but alike in other respects. The change from 3·75 to 4·55 flattens the maximum and moves it from $x = 3 \cdot 0$ to $2 \cdot 6$.

(12.2.1.3) *Ratio of Conductor and Dielectric Losses* (α_d/α_c).—The curved-bracket term also depends upon α_d/α_c. This ratio

can be varied in three ways for a cable of a given size, by variations of g, power factor and frequency. The variation of g has been considered already. The effect of power factor can be seen from curves (*d*), (*b*) and (*e*) of Fig. 6, which show the power rating of the medium-size lead-sheathed cable for power factors of 0·0000, 0·0004 and 0·0020, respectively. It will be seen that an increase in power factor flattens the curve a little and moves the maximum to lower values of x. However, the changes due to power factor are not very marked and can be neglected in most cases. The ratio α_c/α_d varies as $1/\sqrt{f}$. In the case of the cable to which curve (*b*) refers, an increase in frequency from 1000 to 25000 Mc/s would produce a curve of the shape (but not position) of curve (*e*), and a reduction of the frequency to zero would give the shape of curve (*d*). From this we see that the frequency dependence of the curve shape is small.

(12.2.1.4) *Ratio of Depth Underground to Cable Diameter* (l/D_s).—The dependence of the shape upon the ratio l/D_s is shown by curves (*b*) and (*c*), for which $4l/D_s = 180$ and 72. Curve (*c*) corresponds to a series of cables for which D and D_s are 0·800 in and 1·00 in (as for Uniradio No. 10). A comparison of these curves shows that, while the power rating of the larger cable is greater almost in proportion to its increase in size, the shape is not greatly influenced by the change in the ratio of the depth of burial to the diameter of 180 : 72.

(12.2.2) Lead-sheathed Cables installed in Air.

If a cable is installed in air the external thermal resistance per unit length can be taken as

$$S_a = \frac{1}{\pi D_s \theta_s^{\frac{1}{4}} k}$$

where θ_s = Temperature of the outer sheath.
k = Constant allowing for the emissivity of the cable sheath.

In eqn. (13), S_a replaces S_g and gives the power rating for cables installed in air.

$$P_\theta = \theta_c \frac{B}{D} H_g(x) \left(\frac{G_d}{2\pi} \log x \frac{\frac{x}{g}\frac{\alpha_c}{\alpha_d} + \frac{1}{2}}{\frac{\alpha_c}{\alpha_d} + 1} + \frac{1}{\pi D_s \theta_s^{\frac{1}{4}} k} \right) \text{watt} \quad . \quad (15)$$

As the sheath temperature θ_s depends upon the heat generated within the cable, this is also a function of x, and so the second term in the brackets is not constant as it was in eqns. (13) and (14). This has a slight effect upon the shape of the curves. Curve (*f*) in Fig. 6 shows P_θ plotted against x for the same cable as for curve (*c*), except that it is installed in air. It will be noted that installing the cable in air instead of in the ground reduces the power rating by about 25%, but it has little effect upon the shape or value of x_{opt}.

If the temperature of the inner conductor is raised, so is θ_s. The curve (*g*) in Fig. 6 is for the same range of cables as curve (*f*) but with the inner conductor run at 200° C instead of 85° C. (In order to plot the curve on the same decade of the graph, $P_\theta/10$ has been plotted instead of P_θ.) It will be seen that the increase in θ_c has not greatly altered the shape of the curve. The two curves (*f*) and (*g*) are of practical interest, since they refer to a pair of cables of the Uniradio No. 10 size with dielectrics of polythene and polytetrafluorethylene respectively. They show that conclusions drawn about the optimum design of cables with polythene dielectrics will also apply to p.t.f.e. cables.

(12.2.3) Cables with Braided Outer Conductors and P.V.C. Sheaths.

So far as power rating is concerned these cables differ from the lead-sheathed cables in two respects. First they have braided conductors for which g may be between 2·0 and 4·0, and secondly, in general, they have a p.v.c. protective sheath over the braid. For a braided cable of diameter over dielectric of about 0·4 in, g would be 3·5 and the braid would act in the same way as the lead sheath. The only differences in power-rating properties would be due to the thermal resistance of the p.v.c. sheath. This thermal resistance S_c will be added to the other thermal resistances of eqn. (13). S_c will be a constant term and so will have the effect of flattening the curve; also, as it will in general be smaller than the other two terms, its effect will not be great. (For a 50-ohm cable of the 7·25 mm size, such as Uniradio No. 67, RG-8/U or KX50MD1, the thermal resistances will be of the order of $S_d = 88$, $S_c = 48$ and $S_a = 200$ thermal ohms.)

For this reason the curves in Fig. 7, which were calculated for buried cables without p.v.c. sheaths, can be used to represent the general shape of the corresponding curves for p.v.c.-covered cables. The error involved has been calculated for the $g = 1$ curve assuming $S_c = 48$ thermal ohms. Such a thermal resistance would change x_{opt} from $2{\cdot}2_7$ to $2{\cdot}4_2$ and Z_0 for the corresponding solid polythene cables from 33 to 35 ohms. These changes can be neglected in this study.

(12.3) The Optimum Values of Diameter Ratio for Criteria other than that of Cable Diameters

The optimum values of x deduced when the cable is judged by criteria other than that of cable diameter are, in general, little affected by the change in criterion. This is the justification for concentrating solely upon the criterion of cable diameters in the main part of the paper.

(12.3.1) The Criterion of Dielectric Cross-Section.

Where expensive dielectrics such as polytetrafluorethylene are used, it is important to know that efficient use is being made of the material. The area of dielectric cross-section of the cable is

$$A_d = \frac{\pi}{4}(D^2 - d^2) = \frac{\pi D^2}{4}\left(1 - \frac{1}{x^2}\right) \quad . \quad . \quad (16)$$

Hence, from eqn. (2), for a given conductor attenuation α_c, the dielectric cross-section is given by

$$A_d = \frac{\text{constant}}{\alpha_c^2}\left(1 - \frac{1}{x^2}\right)[H_g(x)]^2 \quad . \quad . \quad . \quad (17)$$

The two curves in Fig. 8 show this variation of dielectric cross-section with change in x for constant attenuation, for cables constructed with $g = 1{\cdot}0$ and 3·5. It will be seen that the curves have minima at $x = 3{\cdot}05$ and 4·9 respectively. The corresponding minima, taking cable diameter as the criterion, are at 3·59 and 5·2, but as the curves are so flat the extra cross-section at $x = 3{\cdot}59$ is only 2% greater than at $x = 3{\cdot}05$. (For 4·9 and 5·2 the difference is even less.)

As these two curves relate to cables of the same attenuation, their spacing shows the importance of a low-resistivity outer conductor in conserving p.t.f.e. dielectric.

(12.3.2) Criteria of Conductor Cross-Section and Cable Weight.

The optimum proportioning to secure minimum conductor cross-section can only be studied for specific conductor constructions. A simple assumption to make is that the inner conductor is solid and that the outer is of thickness proportional to its diameter. In this case the conductor cross-section for conductor loss α_c is given by

$$A_c = \pi\left(\frac{d^2}{4} + hD^2\right) + \pi D^2\left(\frac{1}{4x^2} + h\right) \quad . \quad . \quad (18)$$

where hD is the thickness of the outer conductor.

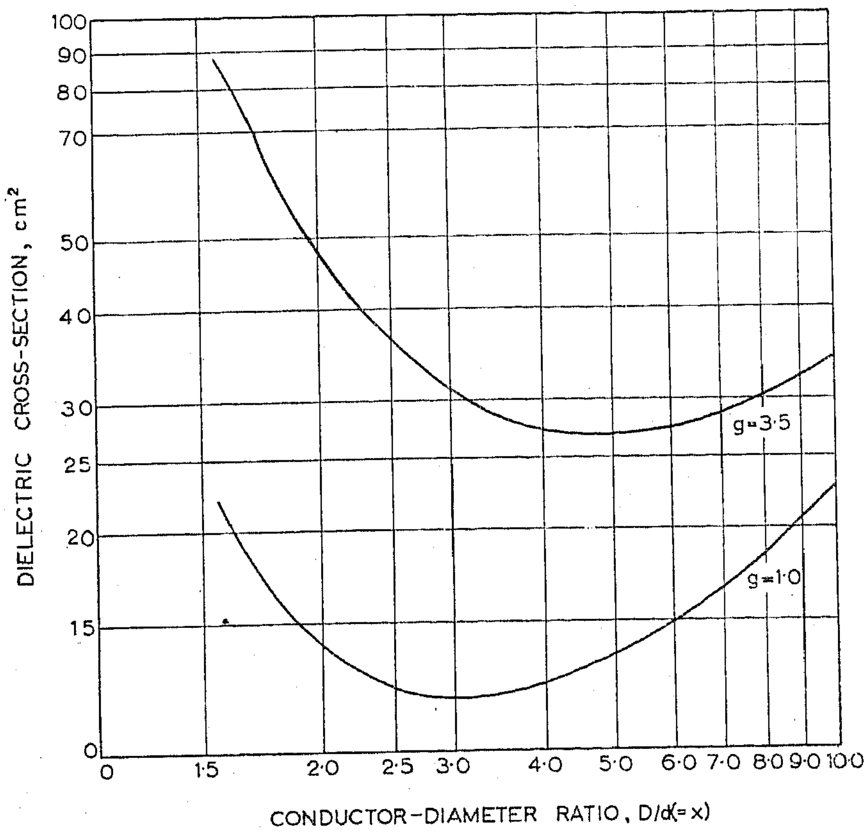


Fig. 8.—Variation of dielectric cross-section with D/d for a given attenuation. The curves are for $\pi A^2/4\alpha_c = 1$.

Hence from eqn. (2)

$$A_c = \frac{\text{constant}}{\alpha_c^2}\left(\frac{1}{4x^2} + h\right)[H_g(x)]^2 \quad . \quad . \quad . \quad (19)$$

Fig. 9 shows the conductor cross-section plotted against x for a cable in which the thickness of the outer conductor is taken as $0 \cdot 14D$. It will be seen that the minimum is at $x = 4 \cdot 3$ and that the excess copper used at $x = 3 \cdot 59$ would be about 1%.

If it is assumed that both conductors are tubes and that as x varies the thicknesses of the tubes are kept constant, then if the thickness ratio of outer to inner tube is b, the conductor cross-section is given by the equation

$$A_c = \alpha\pi d + \alpha b\pi D = \pi\alpha D\left(\frac{1}{x} + b\right) \quad . \quad . \quad (20)$$

Hence from eqn. (2)

$$A_c = \frac{\text{constant}}{\alpha_c}\left(\frac{1}{x} + b\right)[H_g(x)] \quad . \quad . \quad . \quad (21)$$

The Table below shows values of x_{opt} for values of b, and the percentage saving in copper if these values of x_{opt} are used instead of $x = 3 \cdot 59$.

b	2·0	1·0*	0·5	0·2
x_{opt}	4·2	4·7	5·5	8·0
Saving	1%	3%	6½%	9½%

* For $b = 1 \cdot 0$, Hansell and Carter[5] quote without derivation a value of 4·68 for the optimum value.

A cable with a thin-tube outer conductor needs some form of armouring, and the effect of the addition of this on the weight or cost of the cable is the same as that of adding an extra copper sheet of thickness related to the thickness of armouring by some appropriate density or cost factor. As the effective thickness increases, x_{opt} comes nearer to 3·59. It is only in exceptional cases that there would be any appreciable net loss in taking x as 3·59.

The proportioning of a coaxial cable formed from tubes has been considered by Shiniberov.[7] On the assumption that their thicknesses are small compared with the depth of penetration, x_{opt} is given by $bx(\log_\epsilon x - 1) = 1$, for minimum attenuation within a given diameter. The above assumption would be valid only at very low frequencies.

Another form of construction is that in which the inner conductor is solid and the outer conductor is a tube of fixed thickness. Some special cases of this have been studied by Racovitch[6] in connection with the design of carrier cables. A more general analysis shows that x_{opt} varies with the diameter/thickness ratio (D/t) of the outer conductor. For $D/t = 10$ to 50, x_{opt} varies almost linearly from 4·8 to 7·1. For the case of $D/t = 30$, $x_{opt} = 6 \cdot 1$ and the copper saving would be 18% of that needed for a cable of equal attenuation but with $x = 3 \cdot 59$. However, the cable having $x = 6 \cdot 1$ would be larger and require more armour than the cable with $x = 3 \cdot 59$, and the resultant saving in weight would only be 4½%.

It will be noticed that in each of eqns. (17), (19) and (21) the variable term is in two parts, one of which multiplies the $H_g(x)$ term. It is only through this first term that x_{opt} can be displaced from the value calculated for the criterion of cable diameter, which depends upon $H_g(x)$ alone.

In each of the cases considered it has been shown that, while slight gains in cross-sections of dielectric and conductor or overall weight could be achieved in departing from the values of x_{opt} calculated using the criterion of cable diameter, the adoption of this criterion is not likely to lead to any serious error.

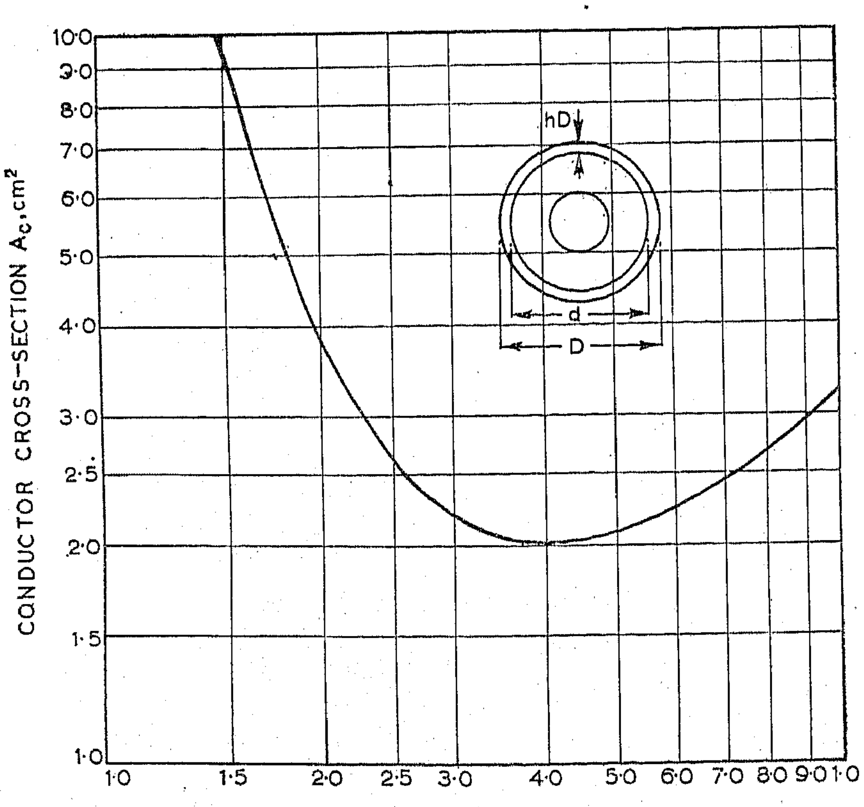


Fig. 9.—Variation of conductor cross-section with D/d for an attenuation of 1 dB/100 ft at 1 000 Mc/s for a line having a solid inner conductor and outer conductor of thickness proportional to its diameter.

(12.4) The Optimum Ratio when More than a Single Property is Specified

In this Section is described a graphical method of determining the optimum diameter ratio when more than one property of the cable is specified. Using this method we could, for instance, determine the diameter ratio which gives the smallest cable to withstand a given voltage and specified attenuation per unit length—or the optimum diameter ratio for a $\lambda/4$ stub which is to have a given input impedance and voltage rating.

Curve (*c*) in Fig. 3 divides the Figure into two parts. To each point in the plane corresponds a cable with diameter and conductor diameter ratio determined by the co-ordinates of the point. All the points in the upper part relate to cables with copper inner and lead outer conductors with α_c less than 1 dB/100 ft at 1 000 Mc/s for $\epsilon = 1 \cdot 0$. If it is specified that a cable is to have this attenuation under these conditions the cable is chosen corresponding to A, the minimum point on the curve, since this has the minimum diameter to meet the specification of attenuation. Let us suppose that, in addition, it were required that the cable should withstand a voltage of $0 \cdot 64$ kV (peak)

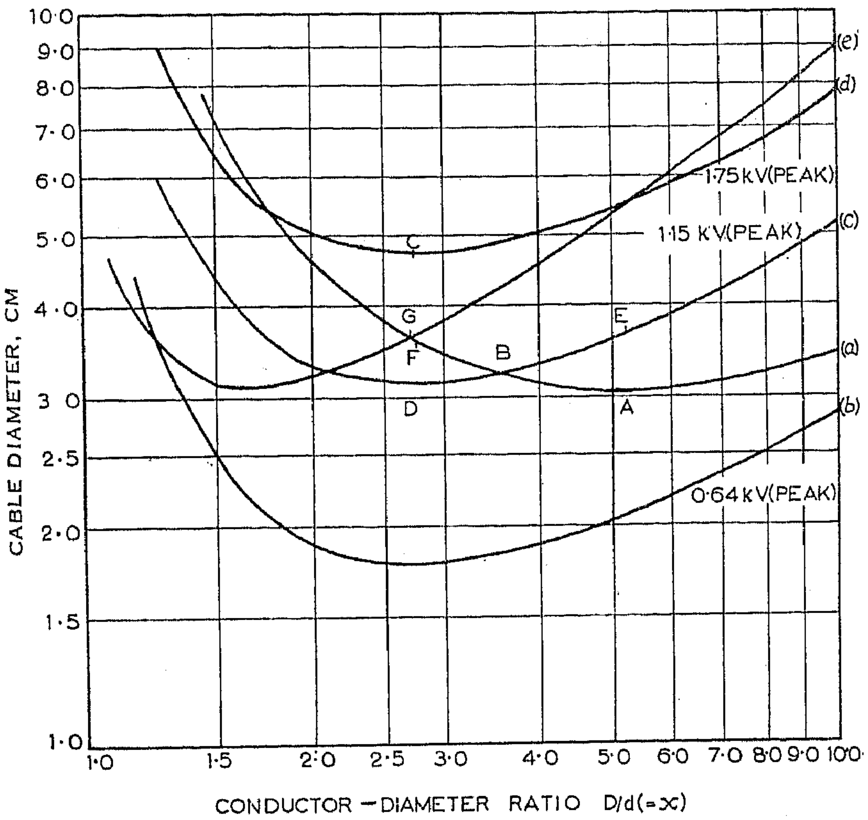


Fig. 10.—Curves for a multiple specification.

assuming $V/E = 0 \cdot 32$ cm and $E = 2$ kV (peak)/cm. Then replotting curve (*c*) of Fig. 3 and the curve of Fig. 4 as shown* in Fig. 10, it is seen that the point A is above the voltage-rating curve for $0 \cdot 64$ kV (peak). The cable corresponding to point A will not only meet the attenuation specification but that of voltage rating as well. However, if the specified voltage rating were $1 \cdot 15$ kV (peak), for which the outside-diameter/conductor-diameter ratio relationship is given by curve (*c*) of Fig. 10, point A would not satisfy the voltage specification. The smallest cable diameter for which both specifications could be met is that corresponding to the point B, the intersection of curves (*a*) and (*c*). The conductor-diameter ratio corresponding to the point B is $3 \cdot 73$, which differs from the optimum deduced from the consideration of either property alone.

As the specified voltage rating is increased from $0 \cdot 64$ to $1 \cdot 15$ kV (peak) it becomes necessary to consider this property as well as the attenuation when choosing the conductor-diameter ratio. It will be seen that it is necessary to consider both properties because their corresponding curves (*a*) and (*b*) have their minima on either side of their intersection. If the specified voltage rating were increased still further to $1 \cdot 75$ kV (peak) as represented by curve (*d*), the intersection of the curves would be outside their minimum values and the smallest cable to meet the specification would be chosen from consideration of voltage rating alone, i.e. it would correspond to the point C.

If, where the specified voltage rating is $1 \cdot 15$ kV (peak) the conductor-diameter ratio were chosen from consideration of attenuation alone, the minimum cable diameter which meets the voltage rating specification would correspond to the point E. This diameter would be $13\frac{1}{2}$% greater than that corresponding to B. If the conductor-diameter ratio be chosen from a consideration of voltage rating alone, the cable would correspond to point F. By a coincidence the increase in diameter in this case would also be $13\frac{1}{2}$%.

Where the cable is also required to transmit an r.f. power such that $\sqrt{P_v}/E = 1 \cdot 0$, the r.f. power curve is obtained by multiplying that of Fig. 5 by a factor of $1 \cdot 33$. This gives curve (*e*) in Fig. 10. It is evident that in this case the minimum cable diameter in which the triple specification can be met is that corresponding to point G, since this is the lowest point on or above the curves (*a*), (*c*) and (*e*).

In the general case the minimum diameter of cable to satisfy a specification of α' dB/100 ft at 1 000 Mc/s, a voltage rating of V′ volts (peak) and an r.f. power (voltage) rating of P'_v for a dielectric of strength E volt/cm and permittivity ε, will be determined by plotting three curves and studying their points of intersection. These curves can be obtained by multiplying the appropriate curve in Fig. 3 by $\sqrt{\epsilon}/\alpha'$, the curve in Fig. 4 by $2V/E$ and the curve in Fig. 5 by $4\sqrt{P_v}/3E\sqrt[4]{\varepsilon}$ or by reference to eqns. (2), (5) and (7).

* Curve (*b*) of Fig. 10 is derived from Fig. 4 by moving it down the diameter scale by a factor of $0 \cdot 64 : 1 \cdot 0$.

The simplicity of the process for multiplying the curves is a result of plotting on log scales. If the curves are plotted on tracing paper the multiplication can be done by moving one sheet of paper over the other and keeping the cable-diameter axis in line. This saves replotting for each problem.

621.395.36 : 621.391

Paper No. 1916
Nov. 1955

TRAFFIC FLOW IN AN EXPONENTIAL DELAY SYSTEM WITH PRIORITY CATEGORIES

By R. E. COX, B.Sc.(Eng.), Associate Member.

(*The paper was first received* 28*th April, and in revised form* 17*th June*, 1955.)

SUMMARY

The mathematical solution to the problem of traffic flow in a simple delay system when the distribution of message lengths is exponential was first established by Erlang in 1917 and is now well known. The paper presents an extension to this theory, covering a case where the traffic is divided into priority categories. Expressions are obtained which describe the behaviour of a general system of this kind and which include that of the homogeneous system investigated by Erlang as a special case. This extension was developed to assist in the study of the effect of using priority categories, including the special category of "deferred" traffic, in automatic and semi-automatic tape-relay systems.

LIST OF SYMBOLS

p_t = Probability function.
n = Expected number of messages per unit time.
T = Mean handling time.
λ = Number of parallel co-operative channels.
ω = Occupancy of the system.
a = Proportion of traffic in the highest-priority category.
(j) = A state where the number of messages undergoing transmission, plus the number of highest-category messages queuing, is j.
P_j = Probability of a state (j).
P'_j = That part of P_j where no lower-category messages are queuing.
P''_j = That part of P_j where lower-category messages are also queuing.
P_D = Proportion of messages which suffers delay.
D_a = Aggregate delay in unit time for messages in the highest-priority category.
(x, y) = A priority category where x is the proportion of traffic in higher categories, and y is the proportion of total traffic in higher categories and in the category defined.
D_x = Aggregate delay in unit time for messages in higher categories than (x, y).
D_y = Aggregate delay in unit time for messages in higher categories than (x, y) and in (x, y).
$D(x, y)$ = Aggregate delay in unit time for messages in category (x, y).
$t_A(x, y)$ = Average delay for messages in category (x, y).
$t_D(x, y)$ = Average delay per delayed message in category (x, y).
$t_f(x, y)$ = Time beyond which only a proportion, f, of messages in category (x, y) is delayed.
$P(a)_{>t}$ = Proportion of highest-category messages delayed beyond t.
$P(x, y)_{>t}$ = Proportion of messages in category (x, y) delayed beyond t.
$t, \bar{t}$ = Any chosen time.
j, q = Any integer.
z = An unknown quantity.
$\omega(x, y)$ = Equivalent occupancy of category (x, y).
t_A = Average delay if the traffic were homogeneous.
t_D = Average delay per delayed message with homogeneous traffic.
t_f = Time beyond which only a proportion f of messages would be delayed, with homogeneous traffic.

(1) INTRODUCTION

In a delay (as distinct from a loss) system, a message offered for transmission but finding all the available channels engaged will wait until a free path is available. The delay system is most tractable to mathematical treatment when the time taken by a channel to handle a message varies in such a way that the probability of a message having a handling time t is $p_t dt = \varepsilon^{-t/T}dt/T$. This is the exponential case; its solution for homogeneous traffic (in the sense that each message has an equal right to the use of the first available channel) was established by Erlang[1] in 1917 and is now well known.

The paper is concerned with the general case where the total traffic is divided into priority categories and where definite proportions of the traffic can be expected to be in each such category. The following definitions should be noted:

(*a*) By "channel" is meant any operator, equipment or circuit through which a message must pass.
(*b*) By "co-operative" channels is meant parallel channels so operated that a waiting message is handled as soon as possible by the first to become free of traffic.
(*c*) By "occupancy" is meant the proportion of unit time that the system is engaged. It lies between zero (channels idle) and unity (channels completely congested). This definition can be extended above unity in the interpretation that if the occupancy is x where $x > 1$, $1/x$ of the total traffic would fully congest the system.

It should also be noted that the work which follows is mathematically correct only if a number of conditions are satisfied, and therefore that the reliability of the results is dependent upon the degree to which they are satisfied. These conditions are as follows:

(*a*) The incoming traffic can be expected to arrive at a definite mean rate of messages per unit time, but individual messages arrive at random times, i.e. messages arrive in a manner to which Poisson's law can be applied.*
(*b*) The mean rate of arrival has been sensibly constant over a long enough time for the probability of finding the system in any specified condition to be independent of time, i.e. for it to have reached statistical equilibrium.
(*c*) Definite proportions of the total traffic can be expected to be in each priority category.
(*d*) The distribution of handling times is the same for all categories.
(*e*) A channel becoming disengaged immediately accepts a message of the highest category available, or if none is waiting, accepts the first to arrive, no matter what its category.
(*f*) Messages in each category are handled strictly in the order in which they arrive. This condition is required only in the investigation of the spread of delays.

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
Mr. Cox is at the Telecommunications Directorate, Ministry of Transport and Civil Aviation.

* Poisson's law states that if incidents are expected to occur at a certain mean rate of n per unit time, but can be expected to occur individually at random, the probability of q incidents occurring in unit time is $n^q\varepsilon^{-n}/q!$.

(2) PRELIMINARY CONSIDERATIONS

Consider a system with λ co-operative channels, handling an average of n messages per unit time with an average handling time T. By definition,

$$\omega = nT/\lambda \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

In the exponential case the probability of a message having a handling time between t and $t + dt$ is

$$p_t dt = \varepsilon^{-t/T} dt/T$$

That this is a correct form is shown by the fact that the probability of its having any length is $\int_0^\infty p_t dt = 1$, i.e. certainty, which is correct, and because the average length $\int_0^\infty t p_t dt = T$, which is by definition.

As n messages are expected to arrive in unit time, the probability of any message arriving in any period dt is

$$ndt = \omega\lambda dt/T$$

With a proportion a of total traffic in the highest-priority category, the probability of a highest-category message arriving in a period dt is

$$andt = a\omega\lambda dt/T$$

It will also be necessary to know the probability that a message found to be undergoing handling will leave during a time dt succeeding the instant of observation.

If it is assumed that the message began its passage through the channel at a time $\bar{t}$ before the instant of observation, then

$$\left.\begin{array}{l}\text{The probability that it will}\\ \text{leave between } \bar{t} \text{ and } \bar{t} + dt\end{array}\right\} = \left\{\begin{array}{l}\text{The probability that it has}\\ \text{a length between } \bar{t} \text{ and } \bar{t} + dt\end{array}\right.$$

$$= p_{\bar{t}} dt = \varepsilon^{-\bar{t}/T} dt/T \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

But this is the general probability for all messages which commenced their passage through the channel at a time $\bar{t}$ before the instant of observation. In this case the message is known to be on the channel after a time $\bar{t}$, i.e. to have a length of at least $\bar{t}$. In other words, eqn. (2) must be expressed as a fraction of the probability of the message having a length at least $\bar{t}$, i.e. as a fraction of

$$\int_{\bar{t}}^\infty p_t dt = \varepsilon^{-\bar{t}/T}$$

The corrected probability is thus dt/T.

As it is independent of $\bar{t}$, it must also represent the probability of a message of any length leaving in a time dt following an instant of observation, i.e. the probability sought.

It follows that if j messages are found to be undergoing handling, the probability of one leaving in a time dt is

$$jdt/T \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

As the system is, by condition, in statistical equilibrium, the probability of finding any specified state of congestion at a time of sampling can be equated to the probability of finding the same state after a subsequent interval dt.

Use of one of the fundamental laws of probability theory then leads to equations of the following type: The probability of a congestion (j) at the end of time dt is equal to the sum of all the permissible products of the form

$$\left.\begin{array}{l}\text{The probability, at the be-}\\ \text{ginning of time } dt \text{, of a}\\ \text{congestion which could}\\ \text{lead to } (j) \text{ after } dt\end{array}\right\} \times \left\{\begin{array}{l}\text{The probability of the oc-}\\ \text{currence, during time } dt,\\ \text{of the step which would}\\ \text{convert that congestion}\\ \text{into } (j)\end{array}\right.$$

The probability of a message joining or leaving the system during dt is an infinitesimal, and the probability of more than one such occurrence during dt is therefore a second-order quantity. Only a small number of initial states can therefore give rise to (j), and the equations in consequence contain only a small number of terms.

(3) BEHAVIOUR OF TRAFFIC IN THE HIGHEST CATEGORY

(3.1) Fundamental Equations

An equation of the type described can now be obtained for every possible state of the system. The full derivation is given below for the most complicated;* the others are derived in a similar manner.

$$P_0 = P_0\left(1 - \frac{\omega\lambda}{T}dt\right) + P_1\frac{dt}{T} \quad . \quad . \quad . \quad . \quad (4)$$

$$P_j = P_{j-1}\frac{\omega\lambda}{T}dt + P_j\left(1 - \frac{\omega\lambda}{T}dt - j\frac{dt}{T}\right) + P_{j+1}(j+1)\frac{dt}{T},$$
$$1 \leqslant j \leqslant \lambda - 2 \quad . \quad . \quad . \quad (5)$$

$$P_{\lambda-1} = P_{\lambda-2}\frac{\omega\lambda}{T}dt + P_{\lambda-1}\left(1 - \frac{\omega\lambda}{T}dt - (\lambda-1)\frac{dt}{T}\right) + P'_\lambda\lambda\frac{dt}{T}$$
$$. \quad . \quad . \quad . \quad (6)$$

$$P_\lambda = P_{\lambda-1}\frac{\omega\lambda}{T}dt + P'_\lambda\left(1 - \frac{a\omega\lambda}{T}dt - \lambda\frac{dt}{T}\right)$$
$$+ P''_\lambda\left(1 - \frac{a\omega\lambda}{T}dt\right) + P_{\lambda+1}\frac{\lambda}{T}dt \quad . \quad (7)$$

$$P_j = P_{j-1}\frac{a\omega\lambda}{T}dt + P_j\left(1 - \frac{a\omega\lambda}{T}dt - \lambda\frac{dt}{T}\right) + P_{j+1}\lambda\frac{dt}{T}, \; j > \lambda \quad . \quad (8)$$

All these probabilities of individual states are independent of time provided that $a\omega \leqslant 1$, although P'_λ and P_j for $j < \lambda$ tend to zero as ω tends to 1 and remain at zero when ω exceeds 1. The above equations are therefore valid so long as $a\omega \leqslant 1$. They simplify to

$$P_1 = \omega\lambda P_0 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

$$(j+1)P_{j+1} = (j + \omega\lambda)P_j - \omega\lambda P_{j-1}, \; 1 \leqslant j \leqslant \lambda - 2 \quad . \quad (10)$$

$$\lambda P'_\lambda = (\omega\lambda + \lambda - 1)P_{\lambda-1} - \omega\lambda P_{\lambda-2} \quad . \quad . \quad (11)$$

$$P_{\lambda+1} = a\omega P_\lambda + P'_\lambda - \omega P_{\lambda-1} \quad . \quad . \quad . \quad (12)$$

$$P_{j+1} - (1 + a\omega)P_j + a\omega P_{j-1} = 0, \; j > \lambda \quad . \quad . \quad (13)$$

Eqn. (10) can be solved successively from $j = 1$ upwards with the aid of eqn. (9) to give

$$P_j = \frac{(\omega\lambda)^j}{j!}P_0, \; j \leqslant \lambda - 1 \quad . \quad . \quad . \quad . \quad (14)$$

Substitution for $P_{\lambda-1}$ and $P_{\lambda-2}$ in eqn. (11) then gives

$$P'_\lambda = \frac{(\omega\lambda)^\lambda}{\lambda!}P_0 \quad . \quad . \quad . \quad . \quad . \quad . \quad (15)$$

Substitution for P'_λ and $P_{\lambda-1}$ in eqn. (12) then gives

$$P_{\lambda+1} = a\omega P_\lambda \quad . \quad . \quad . \quad . \quad . \quad . \quad (16)$$

And the solution of the recurrence relation (13), using eqn. (16), is seen to be

$$P_j = (a\omega)^{j-\lambda}P_\lambda, \; j > \lambda + 1 \quad . \quad . \quad . \quad . \quad (17)$$

* For example, a state (λ) can rise from ($\lambda - 1$) if any message arrives; from (λ) where no lower category messages are queuing, if no highest-category message arrives and no message leaves; from (λ) where lower-category messages are queuing, if no highest-category message arrives [in this case, a message could leave without changing the congestion from (λ), as its place would be taken by a lower-category message]; and from ($\lambda + 1$) if any of the λ messages on the channels leaves. Hence eqn. (7) can be obtained.

(3.2) Determination of the Proportion of Messages Delayed

The proportion of highest-category messages delayed, using eqns. (16) and (17), is

$$P_D = \sum_{\lambda}^{\infty} P_j = [1 + a\omega + (a\omega)^2 + \dots]P_\lambda$$

or

$$P_\lambda = (1 - a\omega)P_D, \quad a\omega \leqslant 1 \quad . \quad . \quad . \quad . \quad (18)$$

P_D is obviously the same for any message and is therefore independent of a. Provided that $\omega \leqslant 1$, a can be set at unity; P_λ then becomes P'_λ. From eqn. (18) using eqn. (15),

$$P_D = \frac{1}{1-\omega}P'_\lambda = \frac{1}{1-\omega}\frac{(\omega\lambda)^\lambda}{\lambda!}P_0, \quad \omega \leqslant 1 \quad . \quad (19)$$

P_0 is then obtained by noting that the sum of the probabilities of all the states of congestion between zero and infinity must be unity, so that, using eqns. (14) and (19),

$$1 = \sum_{0}^{\infty} P_j =$$

$$P_0\left[1 + \omega\lambda + \frac{(\omega\lambda)^2}{2!} + \dots + \frac{(\omega\lambda)^{\lambda-1}}{(\lambda-1)!} + \frac{1}{1-\omega}\frac{(\omega\lambda)^\lambda}{\lambda!}\right]$$

giving* $P_0 = (1 - \omega)\Big/\sum_{0}^{\lambda-1}\left(\frac{\lambda^j}{j!} - \frac{\lambda^{j-1}}{(j-1)!}\right)\omega^j, \quad \omega \leqslant 1$

So that, using eqn. (19),

$$P_D = \frac{(\omega\lambda)^\lambda}{\lambda!}\Big/\sum_{0}^{\lambda-1}\left[\frac{\lambda^j}{j!} - \frac{\lambda^{j-1}}{(j-1)!}\right]\omega^j, \quad \omega \leqslant 1 \quad . \quad (20)$$

For $\omega \geqslant 1$, $a\omega \leqslant 1$, all messages find all the channels busy and are thus delayed, so that

$$P_D = 1, \ \omega \geqslant 1, \ a\omega \leqslant 1 \quad . \quad . \quad . \quad . \quad (21)$$

(3.3) Aggegate Delay in the Highest Category

The aggregate delay in unit time for messages in the highest category is

$$D_a = \sum_{\lambda+1}^{\infty}(j - \lambda)P_j$$

$$= P_\lambda[a\omega + 2(a\omega)^2 + 3(a\omega)^3 + \dots]$$

$$= \frac{a\omega}{1 - a\omega}P_D, \quad a\omega \leqslant 1 \quad . \quad . \quad . \quad . \quad . \quad . \quad (22)$$

(4) BEHAVIOUR OF TRAFFIC IN THE GENERAL CATEGORY

(4.1) Aggregate Delay

Let a general-priority category be denoted by (x, y), where x is the proportion of total traffic in higher categories and y is the proportion both in higher categories and in the category under discussion.

The proportion, x, of higher traffic can be considered as "highest," with the rest, including (x, y), as lower. Or the proportion y of traffic can be considered as "highest" with the rest low. Then, using eqn. (22),

$$D_x = \frac{x\omega}{1 - x\omega}P_D, \quad x\omega \leqslant 1$$

and

$$D_y = \frac{y\omega}{1 - y\omega}P_D, \ y\omega \leqslant 1$$

* Primarily $\sum_{0}^{\lambda}$, but the λ term vanishes.

And if $D(x, y)$ is the aggregate delay of the category (x, y),

$$D(x, y) = D_y - D_x = \frac{\omega(y - x)}{(1 - x\omega)(1 - y\omega)}P_D, \quad y\omega \leqslant 1 \quad . \quad (23)$$

(4.2) Average Delays

The average delay of a category (x, y) is given by

$$t_A(x, y) = D(x, y)\frac{1}{n(y - x)}$$

$$= \frac{T}{\lambda}\frac{1}{(1 - x\omega)(1 - y\omega)}P_D, \quad y\omega \leqslant 1 \quad . \quad (24)$$

The average delay per delayed message in category (x, y) is then

$$t_D(x, y) = \frac{1}{P_D}t_A(x, y)$$

$$= \frac{T}{\lambda}\frac{1}{(1 - x\omega)(1 - y\omega)}, \quad y\omega \leqslant 1 \quad . \quad (25)$$

(4.3) Probability of a Message being Delayed more than a Specified Time

The probability of a message being delayed by more than a specified time can be obtained in respect of the highest-category traffic as follows:

The proportion of highest-category messages which arrive to find j messages entitled and waiting to be served before them is $P_{\lambda+j}$.

Eqn. (3) shows that the probability of a message leaving is then $\lambda dt/T$ for any period dt, and is therefore $\lambda t/T$ for a period t provided that all channels remain busy. Poisson's law then gives the probability of j messages going in time t as

$$\frac{\left(\frac{\lambda}{T}t\right)^j}{j!}\varepsilon^{-\frac{\lambda}{T}t}$$

Therefore the proportion of such messages joining the system at $P_{\lambda+j}$, which have none waiting in front of them at t, is

$$P_{\lambda+j}\frac{\left(\frac{\lambda}{T}t\right)^j}{j!}\varepsilon^{-\frac{\lambda}{T}t}$$

Therefore the proportion of all messages which have none waiting in front of them at a time t after joining the system is

$$\sum_{=0}^{\infty} P_{\lambda+j}\frac{\left(\frac{\lambda}{T}t\right)^j}{j!}\varepsilon^{-\frac{\lambda}{T}t}$$

$$= P_\lambda\varepsilon^{-\frac{\lambda}{T}t(1-a\omega)}$$

The probability of a message going in a following period dt is

$$\frac{\lambda}{T}dt$$

Therefore the proportion of messages which commences transmission between time t and $t + dt$ is

$$P_\lambda\frac{\lambda}{T}\varepsilon^{-\frac{\lambda}{T}t(1-a\omega)}dt$$

And therefore the proportion of messages delayed more than $\bar{t}$ is

$$- P_\lambda \frac{\lambda}{T} \int_{\bar{t}}^{\infty} \varepsilon^{-\frac{\lambda}{T}t(1-a\omega)} dt$$

$$= P_D \varepsilon^{-\frac{\lambda}{T}\bar{t}(1-a\omega)}$$

So that the proportion of highest-category traffic delayed by more than time t is

$$P(a)_{>t} = P_D \varepsilon^{-\frac{\lambda}{T}t(1-a\omega)}, \quad a\omega \leqslant 1 \quad . \quad . \quad . \quad (26)$$

For categories other than the highest, such a direct approach is difficult. It can be seen, however, that all the elemental probabilities of the type "probability of j messages going" or "probability of j messages coming" are exponential with t, and this indicates that the probability of a message in the category (x, y) being delayed by more than time t is also exponential with t.

Therefore let $\qquad P(x, y)_{>t} = P_D \varepsilon^{-\frac{\lambda}{T}tz}$

Then the proportion of messages in category (x, y) delayed between time t and $t + dt$ is

$$-\frac{\partial P(x, y)_{>t}}{\partial t} dt = P_D \frac{\lambda}{T} z \varepsilon^{-\frac{\lambda}{T}tz} dt$$

This represents $n(y - x) t P_D \frac{\lambda}{T} z \varepsilon^{-\frac{\lambda}{T}tz} dt$ of aggregate delay.

Therefore the total aggregate delay of messages in category (x, y) is

$$D(x, y) = n(y - x) P_D \frac{\lambda}{T} z \int_0^{\infty} t \varepsilon^{-\frac{\lambda}{T}tz} dt$$

$$= \frac{n(y - x)}{\lambda z} P_D T = \frac{\omega(y - x)}{z} P_D$$

But from eqn. (23),

$$D(x, y) = \frac{\omega(y - x)}{(1 - x\omega)(1 - y\omega)} P_D$$

So that $\qquad z = (1 - x\omega)(1 - y\omega), \; y\omega \leqslant 1$

and $\qquad P(x, y)_{>t} = P_D \varepsilon^{-\frac{\lambda}{T}t(1-x\omega)(1-y\omega)}, \quad y\omega \leqslant 1 \quad . \quad . \quad (27)$

This form includes eqn. (26)

From eqn. (25) it is seen that

$$P(x, y)_{>t} = P_D \varepsilon^{-t/t_D(x,y)}, \quad y\omega \leqslant 1 \quad . \quad . \quad . \quad (28)$$

(4.4) Time beyond which only a Specified Proportion of Messages is Delayed

If f is the specified proportion of messages and $t_f(x, y)$ is the required time in respect of (x, y), it follows from eqn. (28) that

$$t_f(x, y) = t_D(x, y) \log (P_D/f), \quad y\omega \leqslant 1 \quad . \quad . \quad (29)$$

(5) SUMMARY OF USEFUL EQUATIONS

It may be noted that eqns. (24), (25), (27), (28) and (29) reduce to the corresponding equations for homogeneous traffic if $x = 0$ and $y = 1$, and are in general similar to those equations if $\omega(x, y)$ is substituted for ω, where

$$\omega(x, y) = 1 - (1 - x\omega)(1 - y\omega)$$

$\omega(x, y)$ can be called the "equivalent occupancy" of the category (x, y), and reduces to ω if the traffic is homogeneous.

A list of useful equations can then be rewritten as follows:

$$P_D = \frac{(\omega\lambda)^\lambda}{\lambda!} \Big/ \sum_{j=0}^{\lambda-1} \left[\frac{\lambda^j}{j!} - \frac{\lambda^{j-1}}{(j-1)!}\right] \omega^j, \quad \omega \leqslant 1 \quad . \quad (30)$$

$$= 1, \qquad \omega \geqslant 1 \quad . \quad (31)$$

$$\omega(x, y) = 1 - (1 - x\omega)(1 - y\omega), \quad y\omega \leqslant 1 \quad . \quad . \quad (32)$$

$$t_D(x, y) = \frac{T}{\lambda} \frac{1}{1 - \omega(x, y)}, \quad \omega(x, y) \leqslant 1 \quad . \quad . \quad (33)$$

$$t_A(x, y) = P_D t_D(x, y), \quad \omega(x, y) \leqslant 1 \quad . \quad . \quad . \quad (34)$$

$$t_f(x, y) = t_D(x, y) \log (P_D/f), \quad \omega(x, y) \leqslant 1 \quad . \quad (35)$$

$$P(x, y)_{>t} = P_D \varepsilon^{-t/t_D(x,y)}, \quad \omega(x, y) \leqslant 1 \quad . \quad . \quad (36)$$

One final result is of interest. It can be seen that if t_D, t_A and t_f are the appropriate delays for homogeneous traffic, the effect of introducing a system of priorities is to modify the delays of the category (x, y) in the ratio

$$\frac{t_D(x, y)}{t_D} = \frac{t_A(x, y)}{t_A} = \frac{t_f(x, y)}{t_f} = \frac{1 - \omega}{1 - \omega(x, y)}, \quad \omega < 1 \quad . \quad (37)$$

(6) ACKNOWLEDGMENTS

The author is indebted to Mr. A. J. Bourne of the Ministry of Transport and Civil Aviation for suggesting the problem and for general encouragement, and to Mr. M. H. Johnson, now with Ferranti Ltd., for his comments on an early draft.

(7) REFERENCES

(1) ERLANG, A. K.: "Solution of Some Problems in the Theory of Probabilities in Significance in Automatic Telephone Exchanges," *Post Office Electrical Engineers' Journal*, 1917, **10**, p. 194.

(2) BROCKMEYER, E., HALSTRØM, H. L., and JENSEN, A.: "The Life and Works of A. K. Erlang" (Copenhagen Telephone Company, 1948).

(3) FRY, T. C.: "Probability and its Engineering Uses" (Van Nostrand, 1928).

(4) FELLER, W.: "An Introduction to Probability Theory and its Applications" (John Wiley, 1950), Vol. I.

(5) PALM, C.: "Några Undersökningar över Våntetider vid Telefonanläggningar," *Tekniska Meddelanden från Kungl. Telegrafstyrelsen*, 1937, Nos. 7–9.

(6) PALM, C.: "Specialnummer för Teletrafteknik," *ibid.*, 1946, **104**, p. 1.

Paper No. 1918 R
Nov. 1955

621.317.382.029.64

THE MEASUREMENT OF POWER AT A WAVELENGTH OF 3 CM BY THERMISTORS AND BOLOMETERS

By J. A. LANE, M.Sc., Associate Member.

(*The paper was first received* 12*th April, and in revised form* 1*st July*, 1955.)

SUMMARY

A description is given of measurements made on several thermistors and bolometers at wavelengths of 3·18 and 3·26 cm, using a water calorimeter to investigate their performance as indicators of absolute power. The calorimeter, details of which are given, was designed to measure powers of 2–5 watts; and by means of a 35 dB directional coupler known absolute powers at the milliwatt level were delivered to the thermistor and bolometer mounts. Values of efficiency, defined as the ratio of the power measured in a d.c. calibration to the net input power, of 0·87 to 0·91 were found, with no significant change on varying the operating resistance of the thermistors and bolometers over the range 150–300 ohms.

(1) INTRODUCTION

At radio frequencies of the order of 10^3 Mc/s and above the measurement of power assumes a fundamental importance in radiocommunication, since in this region the basic quantities of current and voltage become difficult to define and measure. At relatively high power levels, above 1 watt, direct calorimetric methods are possible in transmission-line systems, and the water calorimeter has generally been regarded as a standard instrument for the measurement of microwave power. Recently, however, an alternative form of wattmeter has been described by Cullen and Stephenson[1] in which the absolute power flow is derived in terms of measurements of mass, length and time. This instrument has been shown to give readings in good agreement with those of a water calorimeter at power levels in the range 10–30 watts at a wavelength of 3·18 cm (9·44 Gc/s)[2]. For powers in the range 1 μW–10 mW measurements have for some time been made by utilizing the change of resistance with temperature of small semi-conducting beads (thermistors) or fine wires (bolometers).* These are arranged in either coaxial lines or waveguides, so as to absorb the microwave power, and the efficiency of such equipment forms the subject of the present paper.

The bolometer or thermistor is usually connected in a bridge network, by means of which a calibration is conveniently made in terms of direct current or low-frequency power. Errors will normally be associated with this process, however, since the temperature distribution in the resistive element will not be identical in the measuring and calibrating cases. In addition, a small fraction of the microwave power will be dissipated in other parts of the mount, such as the dielectric coating around the element itself, the supporting wires and other metallic surfaces, particularly those in any associated impedance transformers.

There appears to be little quantitative information available in the literature on the magnitude of the errors involved at centimetre wavelengths and the only experimental results known to the author for the 3 and 10 cm wavebands are those of Street[3] and Street and Whitaker.[4] These results show that the d.c. calibration may give errors of a few per cent in these wavebands for typical bolometers and thermistors, the error normally being greater at 3 cm than at 10 cm. A rapid increase in the error as the wavelength is reduced below 3 cm is apparent from the results quoted by Collard, Nicoll and Lines,[5] and by Montgomery[6] for thermistors operated at a wavelength of 1 cm. There seems, however, to be little further information available, especially for bolometers, and the present paper describes a series of measurements of the performance of both thermistors and bolometers at wavelengths of 3·18 cm (9·44 Gc/s) and 3·26 cm (9·21 Gc/s).

* The term "barretter," instead of bolometer, is now common in the United States.

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
The paper is an official communication from the Radio Research Station, Department of Scientific and Industrial Research.

(2) PRINCIPLE OF METHOD

For most practical purposes a knowledge of the magnitude of the individual contributions to the total error mentioned above is not particularly important, and the results obtained in an absolute calibration of any mount may be expressed in terms of an overall efficiency, η, defined by

$$\eta = P'/P'' \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

where P' is the measured power in terms of the d.c. calibration and P'' is the net input power as determined by some standard instrument.

The method of calibration, shown in Fig. 1, was the same, in

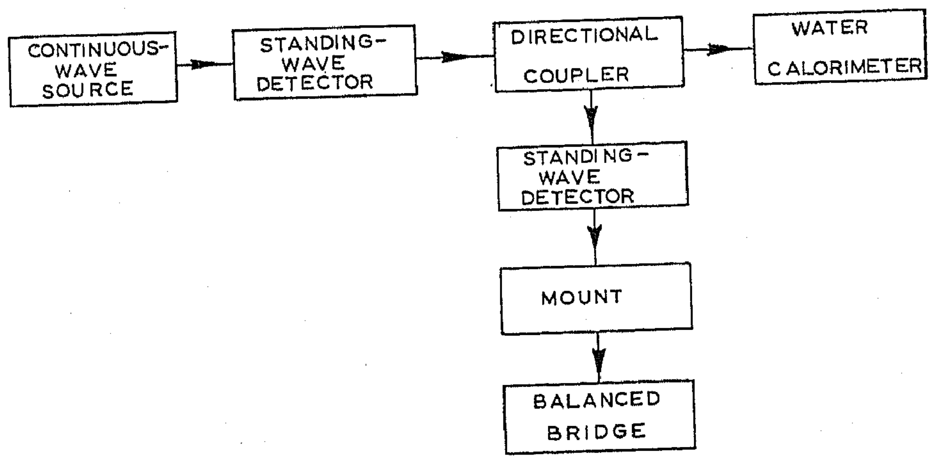


Fig. 1.—Disposition of apparatus for calibration of bolometer and thermistor mounts.

principle, as that described by Street and Whitaker[4] and later by Bailey, French and Lane.[2] Powers of the order of a few watts were measured directly by means of a water calorimeter associated with a calibrated directional coupler. By this technique known powers in the range 0·5–2 mW were established in the waveguide input to the mounts; and these powers were compared with those measured by the bolometer or thermistor when connected in a simple d.c. balanced bridge.

Subsidiary experiments were required to determine the properties of the directional coupler and the calorimeter. For convenience, two precision standing-wave detectors were permanently connected in the calibration equipment so that

measurements could easily be made of the reflection coefficient of the calorimeter and the various mounts under the conditions of calibration.

(3) EXPERIMENTAL EQUIPMENT AND PROCEDURE

While still retaining the method of direct calorimetric measurement, an attempt was made to reduce to a minimum some of the errors inevitably associated with this technique. Details of the assembly finally used for measurements in a 3 cm band waveguide of external dimensions 2·54 cm × 1·27 cm are given below.

(3.1) Water Calorimeter

A continuous-wave source was used for the measurements in order to avoid possible errors associated with the effect of a pulsed signal on both the bolometer elements and the crystal detectors in the standing-wave indicators. A continuous-wave magnetron (type VX 8114) proved to be a satisfactory stable source for output powers in the range 1–10 watts. The use of such a magnetron reduced the amount of attenuation necessary in the directional coupler; but, on the other hand, it entailed greater care in the design of the calorimeter in order to avoid appreciable heat losses. The magnetron required the usual control and stabilizing circuit in conjunction with the high-voltage power unit, so that the anode current could be set and maintained at any desired value in the range 30–50 mA at an anode voltage of 900 volts.

The main features of the calorimeter are shown in Fig. 2. A

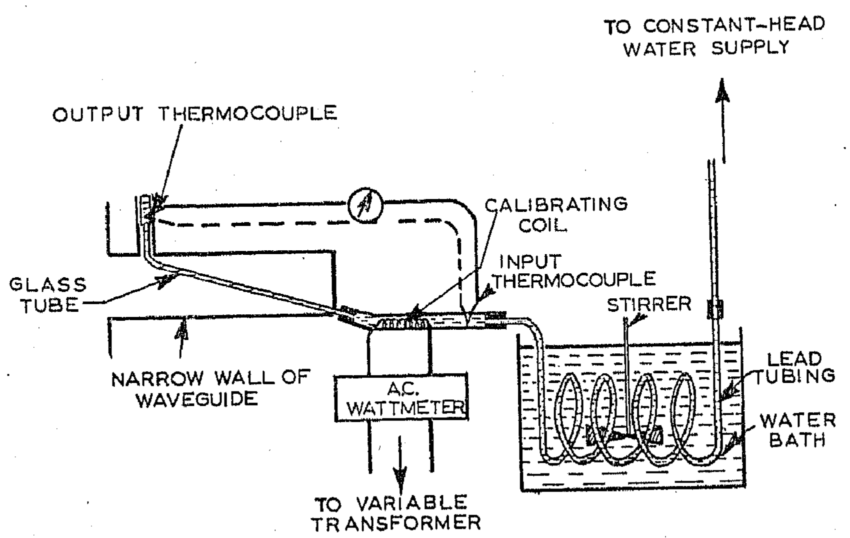


Fig. 2.—Water calorimeter for power measurement at 3·2 cm.

water load consisting of a Pyrex glass tube, 0·3 cm in diameter, extending across the waveguide proved satisfactory for the measurement of powers of a few watts. The reflection coefficient of the load was reduced to a minimum by locating the tube in a plane parallel to the broad face, and the net leakage of power proved to be negligible. A substitution procedure was followed in which the microwave power was measured in terms of the low-frequency power necessary in the calibrating coil to produce the same rise in temperature. For this calibration process a low-frequency (50 c/s) supply was preferred to direct current, since in the latter case electrical leakage through the water causes a current in the thermocouple circuit comparable with that being measured.

The calorimeter was designed for operation at wavelengths in the neighbourhood of 3·2 cm, and its input-voltage standing-wave ratio as a function of wavelength is shown in Fig. 3. This does not represent the best that could be achieved in bandwidth performance, but was regarded as adequate for the present purpose.

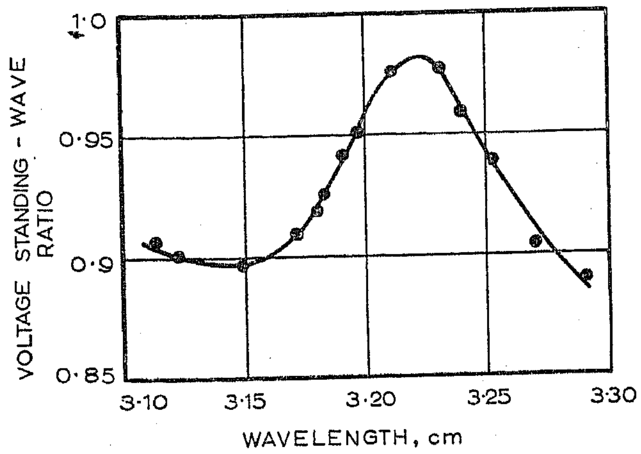


Fig. 3.—Voltage standing-wave ratio of water load as a function of wavelength.

The problem of heat exchange between the calorimeter and its surroundings has already been mentioned. In the present technique an error will arise in the calorimeter measurements, since the heat losses in the measuring and calibrating conditions will not, in general, be equal. These losses were made as small as possible by the following methods:

The whole calorimeter with its inflow and outflow tubes was carefully lagged, first with a double layer of asbestos string and then with closely-packed cotton-wool; in addition, the temperature rise was kept to a minimum compatible with an adequate response in the thermocouple circuit. A rise of 0·5° C, corresponding to a deflection of approximately 25 cm on the galvanometer scale, proved convenient. It was estimated that the normal heat loss between input and output thermocouples was 0·05 watt per degree centigrade rise in the water stream. This corresponds to 0·5% of a power input of 5 watts at a temperature rise of 0·5° C; but in the case of heating by the microwave power the temperature rise is confined to a small portion of the flow system, and the above substitution method will probably tend to overestimate slightly the microwave power in this respect. On the other hand, a small fraction of the microwave power will be absorbed by the glass walls of the load and, although much of the heat so produced will contribute to the temperature rise, some will inevitably be lost from the glass tube. From measurements of the attenuation produced in a waveguide by a similar piece of glass tubing it was estimated that the error due to this cause was unlikely to be as much as 1% and was probably not more than 0·5%.

The a.c. wattmeter was connected so as to read directly the power being dissipated in the heater coil, but a small correction was necessary to allow for power absorbed in the instrument itself. This correction was determined at every measurement and normally amounted to 5% of the total power. The wattmeter reading, after this correction, had an estimated error of not more than 0·2%.

Evidence of the consistency of the observations is given below, but the possibility of some further small concealed systematic error in the calorimeter cannot be entirely dismissed, and it is thought that an error limit of about ±1% should be associated with the mean of any one series of measurements of the high-power level.

(3.2) Directional Coupler

A directional coupler was inserted, as shown in Fig. 1, between source and calorimeter to convey a known small fraction of the output power to the bolometer or thermistor mount. In order to obtain a coupling factor approximately independent of wavelength over a small waveband centred on 3·2 cm, a coupler of the "multi-hole" type was constructed making use of the data

given by Mumford.[7] The dimensions of the coupler, with three holes in the common narrow wall of the waveguide, are given in Fig. 4. When the hole dimensions are so chosen that the relative amplitudes of the component waves admitted from the high-power to low-power arms of the coupler vary from hole to hole in accordance with the coefficients of the binomial expansion, an optimum bandwidth performance is obtained.

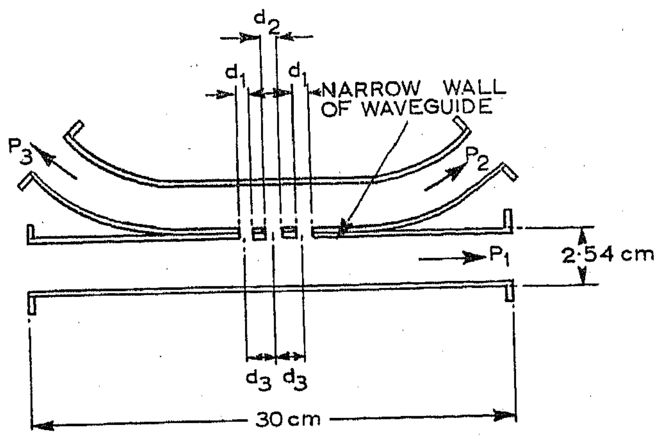


Fig. 4.—Waveguide directional coupler.
$d_1 = 0\cdot595$ cm, $d_2 = 0\cdot714$ cm, $d_3 = 1\cdot12$ cm.

Calculations based on the dimensions shown in Fig. 4 indicated a coupling factor, P_1/P_2, of 33 dB, but as is often found with such instruments, the theoretical value was lower than that found experimentally by 1·5–2·0 dB. The directivity, P_2/P_3, determined experimentally, was 25 dB. One advantage of this type of coupler is that all the waveguide components are mounted in the same horizontal plane, and the amount of mechanical movement involved in a calibration process necessitating the transfer of a detector between low and high-power arms is reduced to a minimum. The calibration itself is described in Section 3.4.

(3.3) Bolometer and Thermistor Mounts and Measuring Bridge

The mounts investigated included two thermistor mounts designed for operation in the waveband 3·1–3·3 cm and a bolometer mount designed for the waveband 3·1–3·6 cm.

The main features of these three mounts are shown in Figs. 5,

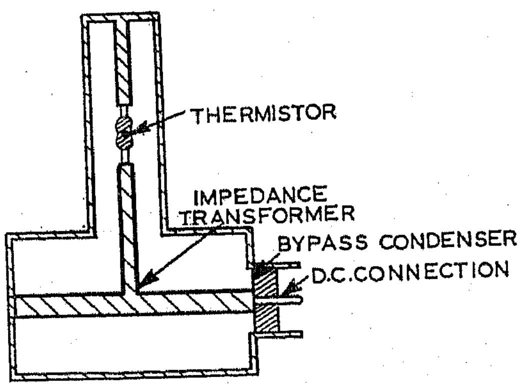


Fig. 5.—3·1–3·3 cm waveband thermistor mount constructed from copper waveguide with a brass coaxial line.

6 and 7. The thermistor mounts are similar in general design, with the thermistor in its glass envelope located in a short section of coaxial line which is joined to the waveguide by means of impedance transformers of the bar-and-post and door-knob types for the mounts of Figs. 5 and 6 respectively. The thermistor bead in each case thus forms part of the inner conductor of the coaxial line and had previously been adjusted in its position in the line to give an optimum-voltage standing-wave ratio in the input waveguide at a wavelength of about 3·2 cm. An interesting feature of the mount shown in Fig. 6 is the

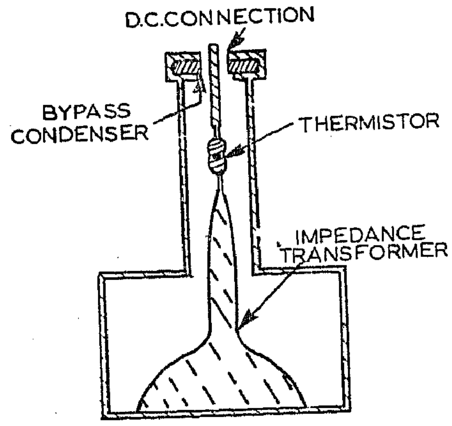


Fig. 6.—3·1–3·3 cm waveband thermistor mount of nickel-plated brass.

conical taper in the coaxial line which transforms the characteristic line impedance from approximately 50 ohms at the point of connection to the waveguide to 200 ohms at the thermistor. The mount shown in Fig. 5 was constructed from copper waveguide with a brass coaxial line, while that in Fig. 6 was constructed throughout in nickel-plated brass.

The bolometer is located directly across the centre of the waveguide perpendicular to the broad face in the mount, the side view of which is shown in Fig. 7, and a resonant iris

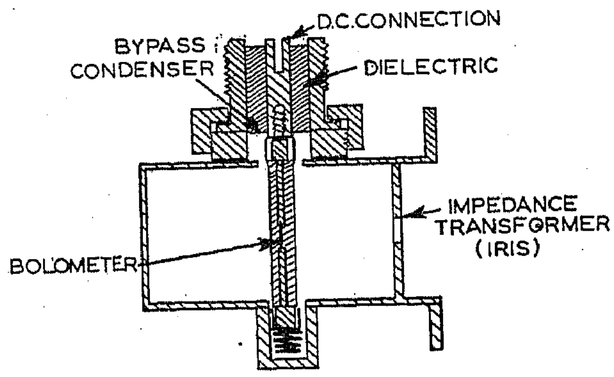


Fig. 7.—3·1–3·6 cm waveband bolometer mount.

impedance transformer in front of the bolometer makes satisfactory performance possible over a broad band of wavelengths. The bolometer element itself consists of platinum wire, 0·25 cm long and $1\cdot5 \times 10^{-4}$ cm in diameter, supported inside a cylindrical capsule. The centre conductor of the fitting is terminated with slotted contact-fingers which hold one end of the bolometer; similar contact-fingers on the opposite face support the other end of the bolometer. The interior of the mount has a silver-plated finish covered with rhodium.

Input-voltage standing-wave ratios for the three mounts are shown in Fig. 8 as a function of wavelength for the resistances

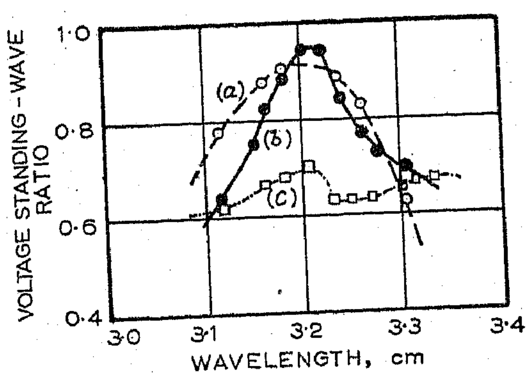


Fig. 8.—Voltage standing-wave ratios of bolometer and thermistor mounts.
(a) Thermistor (Fig. 5) 200 ohms.
(b) Thermistor (Fig. 6) 250 ohms.
(c) Bolometer (Fig. 7) 200 ohms.

shown. The performance of the bolometer mount in the neighbourhood of 3·2 cm is obviously inferior to that of the thermistor mounts, the amount of reflection being about 4% of the incident power.

The bolometers and thermistors were connected in turn in one arm of a simple d.c. bridge circuit as shown in Fig. 9, and

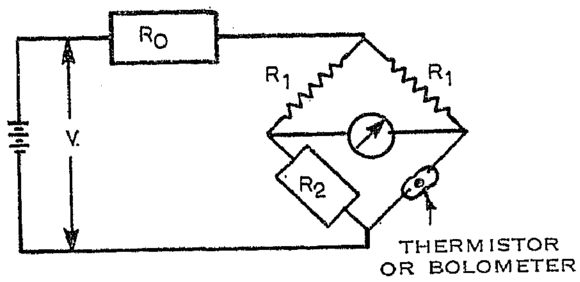


Fig. 9.—Balanced bridge used with thermistor or bolometer.

the d.c. calibration made in the usual way by reducing the bridge current to maintain a balance after the application of microwave power. If the bridge is balanced under d.c. conditions, with the bolometer or thermistor operated at a resistance of R_2 ohms and a resistance R_0 in series with the supply voltage V, and R_0 is increased to $(R_0 + \Delta R_0)$ to maintain a balance with microwave power in the bolometer, then the power in milliwatts may be shown to be given by

$$P_{mW} = \frac{\Delta R_0 V^2 \times 10^3 \times R_2(R + \frac{1}{2}\Delta R_0)}{2R^2(R + \Delta R_0)^2} \quad . \quad . \quad (2)$$

where $R = R_0 + \frac{1}{2}(R_1 + R_2)$. The voltage, V, was measured at each calibration by means of a substandard voltmeter.

(3.4) Calibration of Directional Coupler

The initial observations in this calibration were made in terms of a precision signal-frequency attenuator of the metallized-glass vane type. The setting of the attenuating vane corresponding to the required coupling factor was first determined using a matched crystal detector as a constant-level indicator, and the vane attenuator was then calibrated in terms of an intermediate-frequency piston attenuator in a simple form of superheterodyne receiver. This transfer process was preferred to one in which the receiver was connected directly to the two arms of the coupler in turn, since in this latter process the effect of variations in amplifier gain in the time interval between measurements resulted in a relatively large scatter in the measured values of coupling factor. Care was taken to operate the receiver under conditions which experience had shown gave linear conversion; and matched attenuating pads were included on both sides of the vane attenuator in order to reduce to a minimum the errors due to mismatched components, the significance of which has recently been emphasized by Beatty.[8] The mean result obtained in a series of 17 observations of the coupling factor at a wavelength of 3·18 cm was 34·90 dB, with a standard deviation in the individual observations of 0·08 dB.

As a further check of this initial value, a calibration of the same vane attenuator setting was performed in terms of the balanced bridge shown in Fig. 9 using a thermistor mount, a method which had proved convenient and simple for attenuation values of 10–15 dB. If the setting of the series resistance box corresponds to increases $\Delta R_0'$ and $\Delta R_0''$ above the d.c. value R_0 for microwave power levels of P_x and P_y respectively, it follows that

$$\frac{P_x}{P_y} = \frac{\Delta R_0'}{\Delta R_0''}\,\frac{(1 + \Delta R_0'/2R)}{(1 + \Delta R_0''/2R)}\,\frac{(1 + \Delta R_0''/R)^2}{(1 + \Delta R_0'/R)^2} \quad . \quad . \quad (3)$$

where R as before $= R_0 + \frac{1}{2}(R_1 + R_2)$.

The step of 34·9 dB was too large to measure accurately at one operation, and consequently the vane attenuator setting was calibrated in two stages, each consisting of an attenuation measurement of 17·5 dB. The major difficulty encountered here was the effect of varying ambient temperature when measuring the low-power level, and the assumption was necessarily made that the efficiency of the thermistor and mount remained constant with varying input power. However, the mean value of 18 observations of the coupling factor by this technique was 34·83 dB, with a standard deviation of 0·06 dB in the individual observations.

The uncertainty associated with the above values was still somewhat greater than was desired for the measurements of mount efficiency, and a final calibration was made at the Radar Research Establishment, Malvern, using precision equipment for the determination of microwave attenuation. In this equipment the attenuation to be measured is balanced against the attenuation in a precision piston attenuator operating at an intermediate frequency of 60 Mc/s in a highly-stable superheterodyne receiver. The output meter circuit is so arranged that a discrimination of 0·01 dB is obtained in the measurement of attenuation. The mean of 13 observations of the coupling factor of the directional coupler at a wavelength of 3·18 cm was 34·85 dB, with a standard deviation in the individual observations of 0·02 dB; in a similar series at 3·26 cm the mean value was 34·61 dB. These figures correspond to power ratios of $3{\cdot}05 \times 10^3$ and $2{\cdot}89 \times 10^3$ respectively, with an estimated error limit of not more than 1%, and these were the values subsequently used.

(3.5) Experimental Procedure

The procedure adopted in each calibration was as follows: The bridge network was balanced under d.c. conditions and the magnetron current adjusted to give the required output power. The flow rate through the calorimeter was then adjusted to produce a temperature rise of 0·5° C, as indicated by the galvanometer deflection, and the bridge re-balanced. The system was then allowed to attain a steady state, which required a period of one or two minutes. During this time check measurements were made of the voltage standing-wave ratios in the waveguide input to the calorimeter and the mount. Final adjustments were made on the bridge and the galvanometer deflection was noted; the magnetron was switched off and the low-frequency supply to the calibrating coil switched on simultaneously, its magnitude having previously been adjusted to be approximately the value required. A final adjustment was then made in this supply to reproduce a galvanometer deflection identical with that noted in the measuring condition, this process again requiring a period of one or two minutes. The reading of the wattmeter scale was observed and the appropriate correction made for power absorbed in the instrument.

This sequence of operations was repeated several times for each calibration, and while on occasions efficiencies were found which deviated from the mean by as much as 2%, it was possible in most cases to repeat successive observations so that their maximum deviation was not more than 1%. Although the calibrating coil in the calorimeter had been wound in such a way that it caused an efficient mixing of the water stream, it is still possible that small local temperature gradients existed in the water at the output thermocouple on some occasions or that small air bubbles were causing an abnormal indication. The possibility of a small change with time in the calibration of the mounts themselves cannot be entirely neglected, but an investigation of such an effect, if any, would require experiments of greater accuracy.

The results obtained in a series of observations extending

over a period of about two months are given in the next Section, but it may be mentioned here that they are in general agreement with some subsidiary results obtained using a thermistor bridge calibrated in 1953 against other standard equipment, including the torque-operated vane wattmeter developed by Cullen and Stephenson.[1]

(4) RESULTS AND DISCUSSION

The experimental results are conveniently given, as explained in Section 2, in terms of the overall efficiency of the mount as defined by eqn. (1), where the measured power, P', in the bolometer or thermistor is that given by eqn. (2) and the net input power, P'', is that calculated from the measured calorimeter power using the coupling factor of the directional coupler. A correction, if necessary, was made for the small reflection coefficient of the mounts by means of the data given in Fig. 8. The power reflected P_R is given by the relation

$$P_R = 100\left(\frac{1 - S}{1 + S}\right)^2 \text{ per cent} \quad . \quad . \quad . \quad . \quad (4)$$

where S is the voltage standing-wave ratio.

The mean values obtained for the mount efficiencies at the normal operating resistances are summarized in Table 1, with an estimated error of about $\pm 2\%$.

Table 1

EFFICIENCY OF BOLOMETER AND THERMISTOR MOUNTS AT WAVELENGTHS OF 3·18 AND 3·26 CM

Mount	Operating resistance	Efficiency (η)	
		$\lambda = 3\cdot18$ cm	$\lambda = 3\cdot26$ cm
	ohms		
Thermistor (Fig. 5) ..	200	0·89	0·89
Thermistor (Fig. 6) ..	250	0·90	0·91
Bolometer (Fig. 7) ..	200	0·90	0·91

The errors in the d.c. calibration process are thus approximately 10% for the three types of mount investigated, with no significant difference in performance between the thermistors and bolometers at 3·2 cm. Street and Whitaker[4] have reported a value of η of 0·91 at 3·18 cm for a thermistor bead located inside a ceramic tube in the centre of the waveguide; while for a bolometer in a mount of similar, though not identical, design to that shown in Fig. 7 they found a value of η of 0·96. So far as the author is aware there exist no other precise determinations of thermistor and bolometer performance at 3 cm, although Kerns[9] has reported values of η in the range 0·7–0·95 for various mounts which were deduced as a result of measurements of their input impedance.

It was not possible to insert other bead-type thermistors into the mounts described above and obtain information on the variation in overall performance, but six bolometers of the type described in Section 3.3 were investigated in turn at a wavelength of 3·18 cm in the manner already described, using the mount shown in Fig. 7. For these six bolometers the values of η determined all lay in the range 0·87–0·91.

Some investigations were also made to determine the variation, if any, of η with varying orientation of a given bolometer in its mount. In a series of measurements in which a bolometer was rotated in succession into six positions, the extreme range in η resulting was $\pm 1\%$, with the capsule supported by its contact fingers in the normal way at each position. It was noticed, however, that when the spring contact-fingers on the inner conductor were unscrewed slightly from their normal position against the dielectric by-pass capacitor in the output socket considerable variations in performance were obtained. The standing-wave ratio varied, as the bolometer and its support were rotated, over the range 0·5–0·7, and there were associated variations in η; these values were, for the most part, in the range 0·80–0·92, but values of η as low as 0·70 could be obtained for some positions.

A further series of experiments was performed to determine whether any significant change in efficiency was caused by a change in the d.c. operating resistance. The results are summarized in Table 2 for a wavelength of 3·18 cm.

Table 2

BOLOMETER AND THERMISTOR PERFORMANCE AS A FUNCTION OF D.C. RESISTANCE, $\lambda = 3\cdot18$ CM

Mount	Operating resistance	Range of v.s.w.r.	Efficiency (η)
	ohms		
Thermistor (Fig. 5)	150–300	0·78–0·91	0·87–0·90
Bolometer (Fig. 7)	150–300	0·55–0·70	0·87–0·90

No significant difference from the mean values for a resistance of 200 ohms was found over the range investigated after the correction for power reflection. For very small or very large d.c. resistances an extremely accurate measurement of voltage standing-wave ratio (v.s.w.r.) would obviously be necessary.

(5) CONCLUSIONS

The experiments described above show that errors of 10% may occur in typical thermistor and bolometer mounts at 3·2 cm for power levels of about 1 mW if the d.c. calibration is assumed. For the two bead-type thermistors and six platinum-wire bolometers investigated at 3·18 and 3·26 cm the measured efficiency, defined as the ratio of power measured in a d.c. calibration to the true net input power, was found to be in the range 0·87–0·91. The variation of efficiency with operating resistance is certainly small over the range 150–300 ohms and is probably not more than 2–3%.

At 3·2 cm a water load in the form of a thin Pyrex glass tube placed across the centre of the waveguide in a plane parallel to the broad face has proved satisfactory for power measurements at a level of 2–5 watts, and when associated with a 3-hole-type directional coupler with a 35 dB coupling factor enabled the calibrations described to be performed with an estimated error of $\pm 2\%$.

(6) ACKNOWLEDGMENTS

The author desires to acknowledge the advice given by Dr. J. A. Saxton during the investigation and in the preparation of the paper, and also the help given by Dr. R. A. Bailey of the Radar Research Establishment, Malvern, in the calibration of the directional coupler. The signal-frequency glass-vane attenuator was constructed by Mr. M. W. Barker of the Control Mechanisms and Electronics Division Workshop, National Physical Laboratory.

The work described above was carried out as part of the programme of the Radio Research Board. This paper is published by permission of the Director of Radio Research of the Department of Scientific and Industrial Research.

(7) REFERENCES

(1) CULLEN, A. L., and STEPHENSON, I. M.: "A Torque-operated Wattmeter for 3 cm Microwaves," *Proceedings I.E.E.*, Monograph No. 42 M, July, 1952 (**99**, Part IV, p. 294).

(2) BAILEY, R. A., FRENCH, H. A., and LANE, J. A.: "The Comparison and Calibration of Power-Measuring Equipment at Wavelengths of 3 cm and 10 cm," *ibid.*, Paper No. 1689 R, September, 1954 (**101**, Part III, p. 325).

(3) STREET, R.: "The Absolute Measurement of Low Power at 3000 Mc/s," *ibid.*, Paper No. 838 M, June, 1949 (**96**, Part II, p. 391).

(4) STREET, R., and WHITAKER, P. D.: "Measurement of Microwave Power at Wavelengths of 3 cm and 10 cm," *Proceedings of the Physical Society*, 1950, **63**, p. 623.

(5) COLLARD, J., NICOLL, G. R., and LINES, A. W.: "Discrepancies in the Measurement of Microwave Power at Wavelengths below 3 cm," *ibid.*, p. 215.

(6) MONTGOMERY, C. G.: "Technique of Microwave Measurements" (M.I.T. Radiation Laboratory Series, McGraw-Hill, New York, 1947), p. 154.

(7) MUMFORD, W. W.: "Directional Couplers," *Proceedings of the Institute of Radio Engineers*, 1947, **35**, p. 160.

(8) BEATTY, R. W.: "Mismatch Errors in the Measurement of Ultra-High-Frequency and Microwave Variable Attenuators," *Journal of Research of the National Bureau of Standards*, 1954, **52**, p. 7.

(9) KERNS, D. M.: "Determination of Efficiency of Microwave Bolometer Mounts from Impedance Data," *ibid.*, 1949, **42**, p. 579.

DISCUSSION ON "THE LAUNCHING OF A PLANE SURFACE WAVE"*

Dr. A. L. Cullen (*communicated*): The author derives a very simple formula for launching efficiency for a chopped-surface-wave source. In an earlier paper on this subject† I have studied the same problems and arrived at a different result.

In Fig. A launching efficiency is plotted against k_1h, or kXh in the author's notation. The continuous curve represents the author's formula, and the broken curve is my result, obtained by numerical integration. It is important to know the reason for the discrepancy. In my view it arises because the author has assumed a distribution of tangential electric and magnetic fields in the aperture which is not self-consistent.

The inconsistency can be avoided by assuming a distribution for the tangential electric field only; from this the entire electromagnetic field everywhere to the right of the screen can be evaluated, including the magnetic field "*just* to the right" (by as small a distance as we please) of the aperture plane.

This distribution of magnetic field will be different from that assumed by the author, but it will be consistent with the assumed electric-field distribution. An application of the Poynting-vector approach to a calculation of the power flowing through the aperture due to this magnetic field and the assumed electric field will yield the correct result.

In my calculation the distribution of tangential electric field only was assumed, and the total electromagnetic field was deduced from this. The power radiated was then found by numerical integration of the power polar diagram, instead of calculating the total power flow through the aperture; either result, together with the calculated surface-wave power (about which we are in agreement), will yield the launching efficiency shown by the broken curve in Fig. A.

The author's assumed distribution might be expected to be more nearly valid for large apertures, and the Figure suggests that it is adequate for calculation of launching efficiency in this particular case ($k_1 = 0 \cdot 5k$, or $X = 0 \cdot 5$) when $k_1h > 1$. This criterion of validity will not necessarily hold for other values of X.

The experimental results in Figs. 5 and 6, although valuable in themselves, do not really throw any light on the problem, because the relationship assumed by the author between the field ratio E_r^2/E_a^2 and the launching efficiency is not valid, for reasons outlined above.

* RICH, G. J.: Paper No. 1783 R, March, 1955 (see **102** B, p. 237).
† CULLEN, A. L.: "The Excitation of Plane Surface Waves," *Proceedings I.E.E.*, Monograph No. 93 R, February, 1954 (see **101**, Part IV, p. 225).

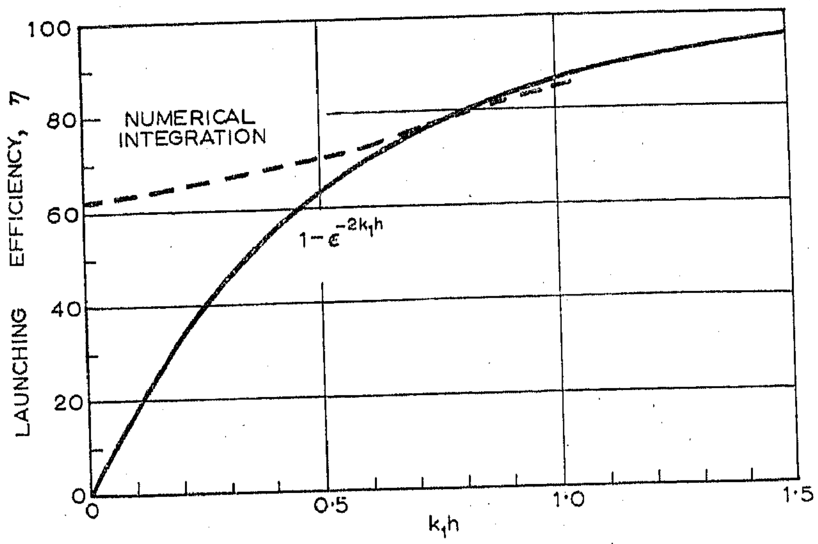


Fig. A.

Mr. G. J. Rich (*in reply*): The distribution I have assumed for the field components in the aperture is such that, apart from vanishing when $y > h$, they are unperturbed from their forms on the open surface. This can be regarded as giving a first approximation to Dr. Cullen's result valid down to kXh (or k_1h) $= 0 \cdot 75$. This result (based on the above graph) corresponds to $h \geqslant 2\lambda/3$ under the conditions both of Dr. Cullen's experimental work and my own.

However, my assumed field does represent exactly in the range $0 \leqslant y \leqslant h$ the field incident upon the aperture from the left. (Dr. Cullen takes the field as originating in the aperture and his analysis applies only to that region lying to the right of the aperture.) Furthermore, Dr. Cullen has defined launching efficiency as "the ratio of power in the surface wave to total power delivered to the slot," although he has, in fact, computed it by integration of the field far to the right of the aperture. My own definition is "ratio of power in the surface wave to total power," and I have sought experimentally to compare the surface-wave power with the input power of the launching device.

It is therefore possible that Dr. Cullen's results and my own can be reconciled by considering the transmission coefficient of the aperture for this particular configuration of the incident field. Let τ be this transmission coefficient, defined as the ratio of the total power in the field to the right of the aperture to the power

given by integrating the Poynting vector of the incident field over the area of the aperture. Let P_I be the incident power then

$$P_I = \rho C_0 \int_0^h \varepsilon^{-2kXy}dy$$

and
$$\tau = \frac{P_R + P_S}{P_I}$$

P_R and P_S are respectively the radiated power and the surface-wave power.

Dr. Cullen's computed launching efficiency is

$$\eta_C = \frac{P_S}{P_R + P_S}$$

whereas I have computed

$$\eta_R = \frac{P_S}{P_I}$$

whence
$$\tau = \eta_R/\eta_C$$

which is accurate to the same order as the launching-efficiency calculations.

An exact experimental verification of the result would present considerable difficulty, but a rough experiment with a plane wave incident upon a slot $3\lambda/2$ broad in a sheet of the non-reflecting substance used for the screen in my experimental work shows that the slot does not present a perfect match to the oncoming field.

Concerning the validity of the results given in Figs. 5 and 6 of my paper I should like to point out that, if the launching efficiency is taken to be P_S/P_I, these serve to confirm my theoretical conclusions as follows:

$$\eta_R = \frac{P_S}{P_I} = \frac{E_r^2\int_0^\infty \varepsilon^{-2kXy}dy}{E_a^2\int_0^h \varepsilon^{-2kXy}dy} = \frac{E_r^2}{E_a^2}\,\frac{1}{1-\varepsilon^{-2kXh}}$$

but the experiment confirms that $E_r^2/E_a^2 = (1 - \varepsilon^{-kXh})^2$, so $\eta_R = 1 - \varepsilon^{-2kXh}$, which is in accordance with theory.

I should like to thank Dr. T. B. A. Senior for several helpful discussions on this topic.

DISCUSSION ON "STANDARD FREQUENCY TRANSMISSIONS"* "THE STANDARD FREQUENCY MONITOR AT THE NATIONAL PHYSICAL LABORATORY"† AND "STANDARD FREQUENCY TRANSMISSION EQUIPMENT AT RUGBY RADIO STATION"‡

NORTH-EASTERN RADIO AND MEASUREMENTS GROUP, AT NEWCASTLE UPON TYNE, 7TH MARCH, 1955

Mr. G. H. Hickling: Could the authors give further information concerning the general policy leading to the setting-up of the various high-accuracy frequency standards now existing in this country? There are at least four of these, that operated by the Greenwich Royal Observatory, which is used to control the B.B.C. and Rugby time signals, the Post Office standard at Dollis Hill, and the MSF and N.P.L. standards. This apparently unnecessary duplication may well be deliberate policy based on the argument that, by means of intercomparisons between different stations (as distinct from comparisons between separate oscillators in the same station, where common disturbing factors may operate), a greater degree of accuracy and reliability can be assured. There would, however, appear to be arguments in favour of locating all equipment in one, or perhaps two, locations, when intercomparisons without the aid of radio transmission would be possible and technical supervision would also be simplified.

It would be interesting to know the reasons which led to the choice of multivibrators for the frequency dividers at MSF, in view of the fact that previously the harmonic-modulator type of circuit has been recommended§ for this duty. An advantage of the latter circuit is that the output voltage ceases in the event of failure of the input signal. Some comment on the steps taken to achieve the apparently high stability of division ratio with the multivibrators would also be useful.

With reference to Fig. 9 of Mr. Steele's paper, it is not evident why recourse was made to an electro-mechanical device for the purpose of making the 25 c/s frequency addition to the 1·5 Mc/s signal. Quite simple electronic circuits can be used to give the same result. In Fig. 11, also, the prolonged large-amplitude frequency deviations require more explanation than is given in the paper. At one period an average deviation of over 2×10^{-7} in frequency persists for about an hour, indicating an integrated time difference of ¾ millisec or more. In terms of a change of path length this would call for an apparently excessive change of about 150 miles. On the other hand, if transitions between E- and F-layer reflections occur abruptly, it seems probable that some cycles of the 2½ Mc/s signal could be lost (or gained) at each transition, causing effectively a "slip" between the transmitter and receiver such as may explain the observed phenomena.

* Essen, L.: Paper No. 1677 R, July, 1954 (see 101, Part III, p. 249).
† Steele, J. McA.: Paper No. 1765 M, October, 1954 (see 102 B, p. 155).
‡ Law, H. B.: Paper No. 1762 R, October, 1954 (see 102 B, p. 166).

§ Booth, C. F., and Laver, F. J. M.: "A Standard of Frequency and its Applications," *Journal I.E.E.*, 1946, 93, Part III, p. 223.

Mr. H. M. S. Smith: It has been said that, by integration over a 24-hour period, the frequency errors (see Figs. 11 and 12 of Mr. Steele's paper) are cancelled out. It would be interesting to learn whether this is, in fact, true, and to obtain a value for the integrated error. It would appear that, so far as remote comparisons by radio links are concerned, the limitation in accuracy is imposed by the errors arising from transmission. It therefore follows that, unless the integrated error due to transmission is smaller than the integrated error due to the oscillators, there is little to be gained in striving for a higher-precision standard, such as an atomic oscillator, unless such standards are used on a purely local basis.

I understand that in this country there are some 150 principal users of the MSF transmissions, and it would seem surprising that such a small number can economically justify the service. If the above argument is valid, it would, in any case, appear that these users will be compelled to use local standards if higher accuracy is required.

Dr. L. Essen (*in reply*): The duplication of frequency standards which Mr. Hickling discusses is, I think, necessary for the work of the different establishments, because of their different functions, although it may have arisen in the first place from the way the quartz rings were developed. They were designed and used at the N.P.L. in 1937, but although one was made for the Royal Greenwich Observatory and two for Australia, the N.P.L. was not equipped for making them in any number. The Post Office therefore performed a very valuable work in further development and in making a number for various different laboratories throughout the world. There are some obvious advantages in geographical distribution, and since all the results are interchanged, the best mean standard can be used as the standard frequency at the N.P.L. and as the standard of time at the Royal Greenwich Observatory. The radio intercomparison introduces no significant error for this purpose. As regards the separate standards at Rugby for the control of the MSF transmissions, it was decided—and the decision has I think been justified—that this was the only satisfactory way of operating the service.

In his comments, Mr. Smith seems to overlook the use of the low-frequency transmissions which were dealt with in some detail in the paper and in the London discussion. The transmissions have shown that frequencies can be compared throughout the world with a precision of one or two parts in 10^9; and therefore give weight to the importance of developing a basic standard of this order of accuracy. I agree, of course, that the use of the high-frequency transmissions is limited to lower accuracies.

Mr. J. McA. Steele (*in reply*): As Mr. Hickling has suggested, purely static methods can be used to produce the same result as the rotating phase-shifter, but whether these deserve to be called "quite simple" is debatable.

In the most usual arrangement, three pairs of sidebands are generated by amplitude modulation of each phase of a 3-phase carrier, the three modulating voltages differing in phase relatively by 120°. On combining linearly the modulated phases cancellation results, not only of the carrier components, but also of all lower (or upper) sidebands, while the remaining upper (or lower) sidebands combine in phase to yield the desired output, differing in frequency from the original carrier by the frequency of the modulation. The correct operation of the circuit depends on a nice balance in both phase and amplitude of the nine component vectors, and in point of ease of adjustment and long-term stability the method is considered inferior to that adopted.

Far from being excessive, the figure of 150 miles quoted for the change in path between F- and E-layer reflection at sunrise is the least that might be expected. If we take the equivalent heights of the E- and F-layers as 110 and 250 km respectively, the difference in the total equivalent paths between Rugby and Teddington for the two modes of transmission is about 260 km or 160 miles. Actual measurements of the reception times of the 1 c/s pulses on MSF 2·5 Mc/s over the sunrise period show a variation of approximately 1 millisec, corresponding to a change in equivalent path of 300 km or 186 miles.

In reply to Mr. Smith, the mean value of the received frequency of MSF 2·5 Mc/s over a 24-hour period, as obtained from records such as that shown in Fig. 11, is found to differ by less than 1 part in 10^8 from the transmitted frequency.

DISCUSSION ON "SOME ASPECTS OF THE DESIGN OF V.H.F. MOBILE RADIO SYSTEMS"*

Mr. K. L. Rao (*India: communicated*): The author has made a statement that major improvements are unlikely in amplitude modulation. The single sideband will offer future possibilities of economy in frequency usage. The chief drawback has been the difficulty of isolating the sideband by simple means. It may be that a system will be developed where only one sideband is produced by modulation methods instead of having resort to filter circuits to achieve this.

The United Kingdom appears to favour amplitude modulation. Perhaps the earlier systems—even television—have been developed on amplitude-modulation systems, in which case it may be a major economic problem to change over to frequency modulation at this stage. Also, as pointed out by Mr. Wheeldon, amplitude-modulation equipment is more easily maintained by semi-skilled personnel; this is a very attractive feature and weighs heavily in favour of amplitude modulation. In so far as the particular v.h.f. mobile system dealt with by the author is concerned, there is nothing much to choose between the two systems of modulation.

Mr. E. P. Fairbairn (*in reply*): Tests would be needed to prove whether in fact the single sideband has any advantages for these systems, where selective sideband fading is not a problem. There is at present no prospect of producing a competitive single-sideband system, since the problems have not been solved even in the h.f. band for mobile applications.

The difficulties of maintenance of frequency-modulation equipment by semi-skilled personnel are entirely due to lack of experience, and will disappear as frequency modulation comes into more general use—say for v.h.f. broadcasting.

* FAIRBAIRN, E. P.: Paper No. 1522 R, July, 1953 (see **101**, Part III, p. 53).

621.396.96.08 : 621.396.11

Paper No. 1917 R
Nov. 1955

DETERMINATION OF THE REFLECTION COEFFICIENT OF THE SEA, FOR RADAR-COVERAGE CALCULATION, BY AN OPTICAL ANALOGY METHOD

By G. W. G. COURT, B.Sc., A.Inst.P., Associate Member.

(*The paper was first received* 10*th February, and in revised form* 8*th June*, 1955.)

SUMMARY

An approach to the solution of the problem of the effect of a rough sea surface on the vertical coverage of a radar set, by means of an optical analogy, is suggested. The reflection of light waves by a ground-glass surface is considered, and the results of optical experiments are applied to the case of radio waves reflected by the surface of the sea. The method does not yield a precise solution to the problem, but the result achieved does not conflict seriously with practical results which have been recorded up to the present.

A typical coverage diagram derived by this method is given to show the effect on the coverage of the variation of sea-surface condition as indicated by the sea-state scale.

(1) INTRODUCTION

In estimating the vertical coverage to be expected from a radar set operating over the sea it is usual to make a number of simplifying assumptions. In particular, the reflecting properties of the sea surface are approximated, so that for long wavelengths, i.e. of the order of metres, the sea surface is regarded as being perfectly smooth, and for very short wavelengths, i.e. the microwave region, it is regarded as perfectly rough. For intermediate wavelengths we have the problem of a semi-rough reflector, and a number of theoretical analyses of the reflection of radio waves from rough and irregular surfaces have been carried out.[1,2] The conclusions reached here are derived by considering an optical analogy in which light waves are reflected by a matt surface. They lead to an approximate solution of the problem of the semi-rough case, and indicate appropriate modification to the diagrams, which are obtained by considering the sea either as perfectly smooth or completely rough.

(2) DERIVATION OF VERTICAL POLAR DIAGRAM IN IDEAL CASE

For a radar set operating at a height h_1 above a reflecting surface (Fig. 1) the method of derivation of the vertical polar diagram,

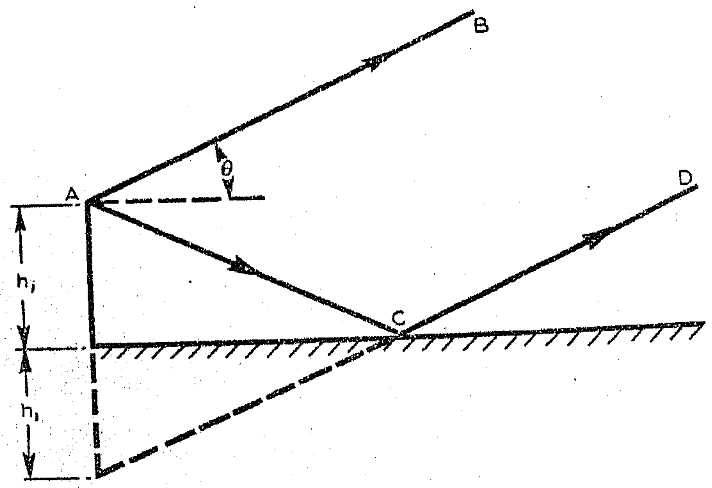


Fig. 1.—Direct and reflected ray paths.

in the ideal cases of perfectly smooth and completely rough surfaces, is well known. The indirect ray ACD can be assumed to be reflected or destroyed at C in the respective cases, and the polar diagram shown in Fig. 2 can be deduced.

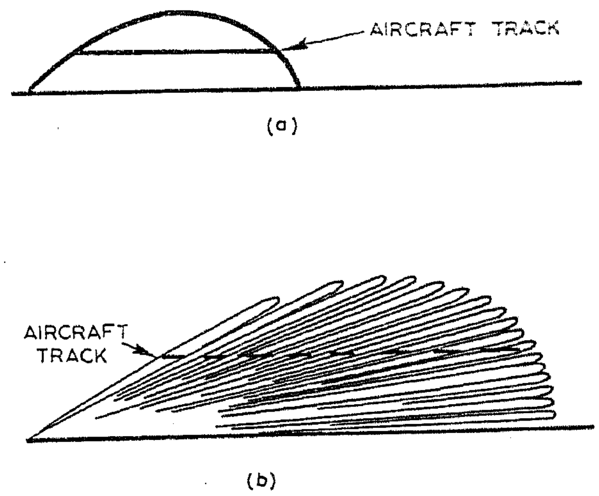


Fig. 2.—Ideal polar diagrams.

(*a*) Free space.
(*b*) Complete interference.

Throughout the paper a polar diagram, or part of a polar diagram, in which no interference effect occurs will be referred to as giving "free space" cover, and Fig. 2(*a*) is labelled "free space" accordingly. Fig. 2(*b*) indicates a complete-interference polar diagram, and in both cases a typical aircraft track is indicated.

(3) REFLECTION OF LIGHT FROM A ROUGH SURFACE

(3.1) Proposed Analogy

Suppose that the surface is semi-rough, but not sufficiently rough to destroy the secondary wavefront. The condition is analogous to that of the reflection of light from a matt surface. Some discussion on this is to be found in well-known textbooks on the subject of light, and it is proposed to apply the results of light experiments and deductions to the radar problem.

(3.2) Discussion in the Literature

In the literature it is stated that the distinctness of an image of a light source reflected in a matt surface depends upon the roughness of the surface and the angle of elevation of the observation. When this angle approaches zero a distinct image is seen, but when it is increased, a critical angle is reached at which specular reflection ceases. Furthermore, as this critical angle is approached the image of a white-light source, which is white at low angles of observation, becomes reddish before it is lost as the critical angle is exceeded. A typical example quoted is a smoked-glass surface, which produces no image at normal incidence but at nearly grazing incidence gives an image of surprising distinctness.

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.

Mr. Court was formerly at the Dominion Physical Laboratory, Department of Scientific and Industrial Research, New Zealand, and is now at Civil Aviation Administration, Air Department, New Zealand.

(3.3) Relation between Critical Angle and Surface Roughness

For a surface having unevenness, as shown in Fig. 3, where a "bump" height is h, a reflected ray has a path length to an observer differing from that for the flat surface by

$$2h \sin \theta$$

where θ is the angle of elevation.

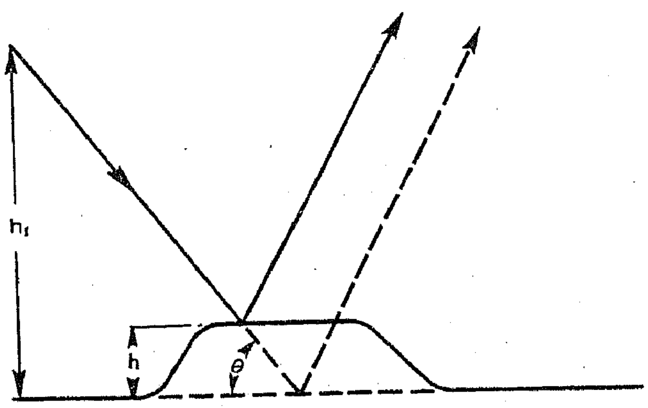


Fig. 3.—Reflection from a rough surface.

Schuster and Nicholson[4] make the definite suggestion that the criterion for specular reflection is

$$h \sin \theta < \lambda/8$$

which corresponds to the Rayleigh criterion.

The criterion will be interpreted in the following manner. The secondary wavefront reflected from a rough surface is coherent and able to interfere with the direct wavefront only if the surface roughness satisfies the condition

$$h < \lambda/8\theta$$

where θ is small so that $\theta = \sin \theta$.

From the expression for path-length difference, $2h \sin \theta$, it is clear that for small values of θ the effect of surface roughness is small and the reflected wavefront at small angles is little affected. If θ is greater than the value given by $h = \lambda/8\theta$, the reflected wavefront is no longer coherent and no image is seen.

(3.4) Variation of Reflection Coefficient with Angle of Elevation

To obtain an indication of the relation between the reflection coefficient and angle of elevation a measurement was made of the change of intensity of the reflected image of a small lamp in a ground-glass surface as the angle of observation was varied. A Macbeth illuminometer was used for this purpose, and the result is shown in Fig. 4. Measurements were satisfactory when the angle of elevation was not too close to grazing incidence, and the relationship appears to be logarithmic.

(3.5) Derivation of Reflection Law

The information available from the optical case may be summarized as follows:

(a) There is a critical angle of elevation θ_c above which specular reflection is not discernible.
(b) The reflection coefficient at $\theta = 0$ is unity.
(c) The law governing the variation of the reflection coefficient with angle of elevation for $\theta < \theta_c$ is logarithmic.

We can write $\log \rho = m\theta$, so that ρ is unity when θ is zero, where ρ is the reflection coefficient and m is a constant.

To determine m we require to establish ρ for a given angle of elevation, and it is convenient to consider θ_c, which has been defined as $\lambda/8h$, and at which specular reflection ceases. Experimentally it was found that measurement of ρ became very

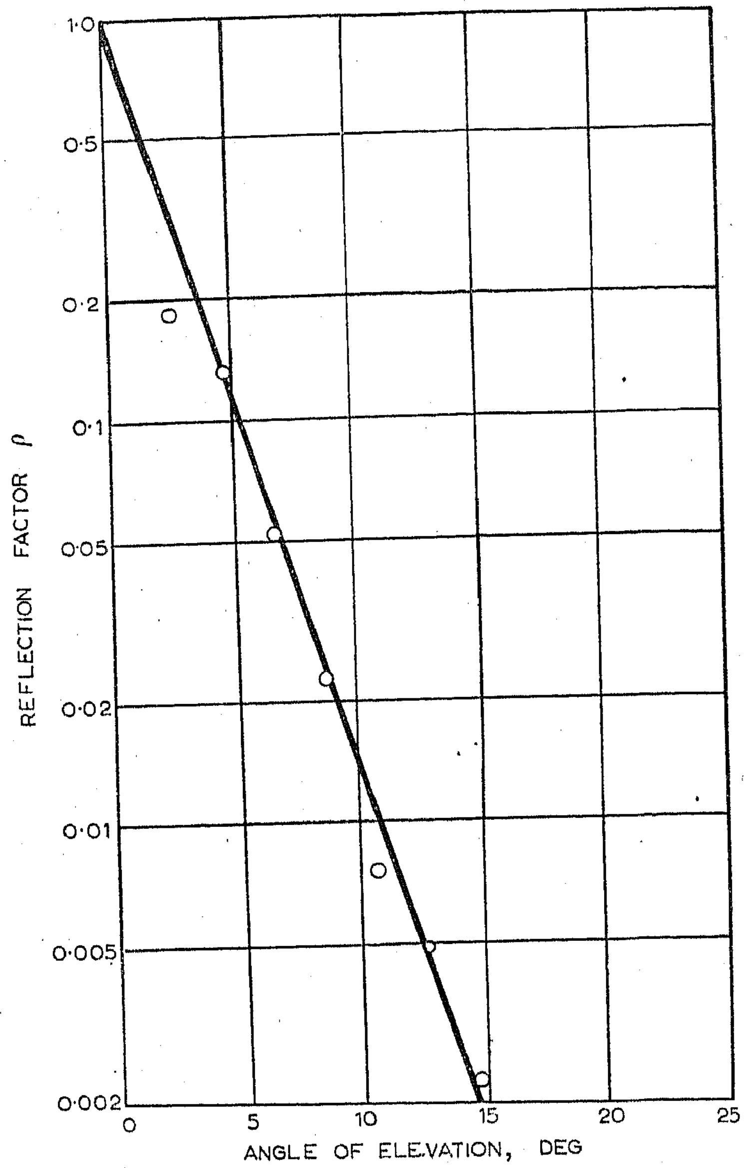


Fig. 4.—Reflection of light from ground-glass surface.

difficult when a value of about $0 \cdot 002$ was reached. It might be considered that at this point specular reflection has ceased.

Then, since $\rho = 0 \cdot 002$ when $\theta = \theta_c$

$$\rho = \exp [(\log 0 \cdot 002)\theta/\theta_c]$$

(3.6) Check of θ_c against Measured Surface Roughness

An attempt was made to compare the practical values of θ_c obtained by the choice of a value for ρ of $0 \cdot 002$ and that derived from a measured value of h substituted in the formula $\theta_c = \lambda/8h$.

The roughness of the ground-glass surface used to determine the relationship between ρ and θ was compared with a roughness standard using a commercial instrument. This indicated an average roughness of the order of 10–20 microns. Taking the wavelength of light as 5 000 Å, the resultant critical angle is 7–14°. This compares favourably with the figure obtained by optical measurements for $\rho = 0 \cdot 002$ (see Fig. 4), i.e. approximately 14°.

As the value obtained for h is an approximation only, the close agreement is considered fortuitous, but it does indicate that the proposed estimate for θ_c should give a result of the right order.

(4) APPLICATION TO THE RADAR PROBLEM

(4.1) Consideration of Light and Radio-Wave Reflection

Considering the sea-surface reflection of radio waves to be analogous to light reflection from a ground-glass surface, the results of the discussion and experiments may be applied to the

problem of a radar-coverage pattern. Although the sea surface is unlikely to be similar in form to the surface of a ground-glass sheet, the optical results should give an indication of the results likely to be obtained for radio waves and the sea surface. Furthermore, although the ground-glass surface is stationary whilst the sea surface is moving, the movement is slow compared with the velocity of electromagnetic waves, and may be expected to produce in the interference pattern a certain amount of scintillation of the same period as the general sea-wave motion. It is significant that in the optical experiments the intensity only of the reflected light is measured, and no consideration has been given to phase changes upon reflection. As a result, two points arise which must be considered when applying any deduced relationship to the radio case.

(4.2) Reflection Factor for Radio Waves

In the light experiment the reflection factor ρ has been determined by comparison of direct and reflected light intensities. In the radio case we require the relationship between direct and reflected radio-wave amplitudes. Thus in the radio case the reflection coefficient will be k, where $k = (\rho)^{1/2}$

i.e. $$k = \exp\left[(a/2)(\theta/\theta_c)\right]$$

where a is a constant, replacing the value $\log 0{\cdot}002$ used in the optical case and determined by the amplitude of the reflected radio wave which is considered to produce the minimum appreciable interference pattern.

Let us assume that the interference pattern can be neglected when the reflected field strength is 5% of the direct field strength. Then at $\theta = \theta_c$, a equals $\log 0{\cdot}0025$, and the appropriate relationship for k is

$$k = \exp\left[\tfrac{1}{2}(\log 0{\cdot}0025)(\theta/\theta_c)\right]$$

(4.3) Modified Form of Radar-Coverage Diagram

If the reflection coefficient k is applied to determination of a radar-coverage diagram, a result as shown in Fig. 5 would be

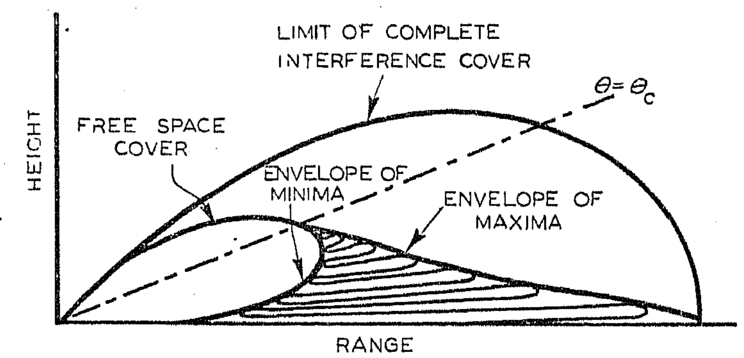


Fig. 5.—Modified polar diagram.

expected. The modification of the original complete interference pattern (Fig. 2) is apparent. There is complete interference at $\theta = 0$, and the amplitude of the interference pattern reduces as θ is increased until it is negligible at $\theta = \theta_c$.

(4.4) Significance of Phase Change upon Reflection

In the radar case, knowledge of the resultant phase of the reflected wave is necessary to derive an interference pattern precisely. However, for a system operating at a short wavelength, so that the site height is great compared with the wavelength, the interference lobes are very closely spaced and their detailed derivation is not generally required. It is sufficient to indicate that, in a particular zone of the coverage diagram, the pattern is a set of closely spaced interference lobes, which is known to show an aircraft movement as a series of short lengths of track, as indicated in Fig. 2. Thus only the envelopes of the maxima and minima of the interference pattern need be calculated. In the zone of the coverage pattern free from interference the track should be a continuous line, as shown in this Figure.

For the sea surface there will be a continuous variation in the phase of the reflected wave, owing to the overall variation of the sea surface. The effect will be to sweep continuously the lobes of any interference pattern over small angles. The pattern as a whole will, in fact, occupy the interference zone, but its detailed structure will vary.

The effect of sea movement on the coverage of a target in the zone is difficult to assess exactly; the general results will be to maintain "interference" cover on the target but to make the resultant track less definitely divided into short lengths. The slight swinging of the lobes which will occur could well be effective in making the track more or less continuous.

(5) PRACTICAL VALUES OF θ_c FOR VARIOUS SEA STATES

It is of interest to consider values of θ_c for different wavelengths and sea-surface conditions. The value of h is taken as the wave height, and this has been interpreted from the scale of sea states as the mean of the wave heights quoted, e.g. for a sea-state scale 4, the wave height is given as between 4 and 8 ft, and h is taken as 6 ft.

The derivation of a value for h in this way is purely arbitrary, but the action is possibly justified in view of the limited information readily available on sea conditions. It is likely that further work on sea-surface conditions may lead to a better interpretation of h. However, the general argument may be followed, and as a result Table 1 can be drawn up, showing the relationship between the critical angle, wavelength and sea state.

Table 1

Sea-surface condition		Critical angle θ_c			
Sea state	h	λ = 50 cm	λ = 25 cm	λ = 10 cm	λ = 3 cm
	ft	deg	deg	deg	deg
1	0·5	23·6	11·8	4·7	1·4
2	1·5	7·8	3·9	1·55	0·47
3	3·0	4·0	2·0	0·8	0·24
4	6·0	2·0	1·0	0·4	0·12
5	10·5	1·12	0·56	0·22	0·07
6	16·5	0·72	0·36	0·14	0·04
7	25·0	0·48	0·24	0·095	0·03

From Table 1 it is to be expected that, even for wavelengths of 10 and 3 cm and under rough sea conditions, an interference pattern should occur at low angles of elevation. Some confirmation of this is given, for sea conditions up to sea state 4, in the results of field-strength measurements off the coast of New Zealand.[3] At longer wavelengths the existence of an interference pattern has been confirmed by standard radar test flights carried out for different sea-surface conditions.

(6) DERIVATION OF COMPLETE COVERAGE DIAGRAM

It is not proposed to describe in detail the calculation of the complete coverage diagram making use of the derived values of k, but an illustration is shown in Fig. 6 of the result of such a calculation. Due account is taken of the earth's curvature. Sea states ranging from 1 to 5 have been considered.

The lobe structure of the interference pattern is not indicated; only the envelopes of the maxima and minima for the various

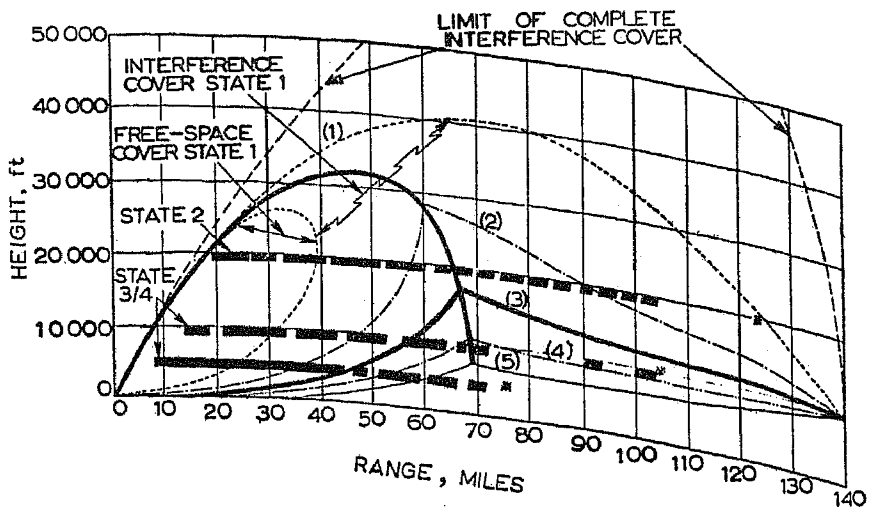


Fig. 6.—Vertical cover for sea states 1–5.
$\lambda = 25$ cm.
*Limit of flight.

sea states are given. Beneath these envelopes an interference pattern exists, the height of the site and the wavelength controlling the structure details, and owing to the sea movements, the structure will be expected to vary somewhat in so far as the lobe positions are concerned. In addition, there is the zone of free-space cover in which no interference occurs. These are indicated, in particular, for a sea state 1.

It is of interest to note that even in very rough sea conditions a considerable increase in range, over free-space range, is to be expected for targets at low angles of elevation.

(7) TEST-FLIGHT RESULTS

A number of aircraft tracks obtained from test flights are shown on the diagram, which was calculated for the appropriate target size and various sea states. The results are insufficient to confirm or refute the estimate of the cover, but from these tracks and a knowledge of other similar test flights, there appears to be some justification for expecting a coverage diagram of the type constructed.

In flight A, for which the reported sea state was 2, the form of track changes at a point near the end of the free-space coverage for a sea state 2, and continues for some distance. It is clear that the range is greater than that to be expected with a completely rough reflecting surface, and that the coverage does not extend to twice the free-space range, as would occur with an interference pattern from a perfectly smooth reflecting surface. The track plotted seems to correspond to a sea state of between 1 and 2. The recorded sea state is estimated, not measured, and an error of unity in the scale would not be unlikely.

In flights B and C, with sea states reported as 3–4, the information is less valuable as the flights were not extended to a sufficient range. Both flights show that, even under choppy sea conditions, the range is extended beyond the free-space range.

(8) VALUE OF FURTHER OPTICAL EXPERIMENTS

Some consideration has been given to more refined optical experiments in which a reflecting surface derived from a photograph of the sea is used. The Cabro printing process could reproduce from a photograph a relief pattern on a gelatine surface. However, it will be found upon examination that the resulting surface does not correspond precisely with the sea surface but, depending on the angle of illumination, more nearly to its first differential. Even so, the periodicity may be reproduced, and this fact can be of value.

However, the information at present available on the sea surface is limited, and the approximate method suggested in the paper is probably sufficient to indicate the radar-coverage pattern to an accuracy comparable with sea-state reporting. There may well be reason to believe that further test flights under various sea conditions would indicate that the method is capable of a higher degree of precision than has so far been established.

(9) CONCLUSION

The use of the optical analogy provides an indication of the vertical coverage to be expected from a radar set operating over a sea surface. It differs appreciably from the polar diagrams to be obtained by considering the sea surface as completely smooth or rough, to an extent which depends on the sea state and radio wavelength used.

It is considered that the arbitrary choice of parameters could be improved if information from an extended series of test flights were available. The choice of a limiting value of $k = 0{\cdot}05$ is possibly somewhat large, and a smaller value may be justified. Such a change would reduce the extent of the interference cover zone. However, the direct application of θ_c as defined optically could be changed, to increase the critical angle in the radio case. Since it may be expected that a radio method may be more sensitive than a visual one in detecting a reflected image, an increase in the limiting angle for a radar interference pattern to a value in excess of $\lambda/8h$ may also be justified. The result of this change would be to increase the extent of the interference cover zone. Results available give the impressions that the present choice of parameters is a reasonable compromise.

(10) ACKNOWLEDGMENTS

The author acknowledges the discussions with several of his colleagues and, in particular, the critical discussions with Mr. W. H. Ward. The help given by Mr. J. Hull in carrying out the illuminometer measurements is also gratefully acknowledged.

This paper is published with the permission of the Secretary of the Department of Scientific and Industrial Research, New Zealand.

(11) REFERENCES

(1) RICE, S. O.: "On Electromagnetic Waves reflected from Rough Surfaces," *Communications on Pure and Applied Mathematics*, 1951, **4**, p. 351.

(2) DAVIES, H.: "The Reflection of Electromagnetic Waves from a Rough Surface," *Proceedings I.E.E.*, Monograph No. 90 R, January, 1954 (**101**, Part IV, p. 209).

(3) Report of Factual Data from the Canterbury Project (D.S.I.R., New Zealand, 1951).

(4) SCHUSTER, A., and NICHOLSON, J. W.: "Theory of Optics" (Edward Arnold, 1924).

621.394.73 : 621.315.28

Paper No. 1908 R
Nov. 1955

SIGNALLING SYSTEMS FOR SUBMARINE TELEGRAPH CIRCUITS

By C. J. HUGHES, B.Sc., Associate Member.

(*The paper was first received* 8*th March, and in revised form* 8*th June*, 1955.)

SUMMARY

An account is given of the characteristics of submarine telegraph cables and terminal apparatus and the noise spectra for simplex- and bridge-duplex-operated cables. By using simplified expressions for the noise spectra it is possible to determine the number of signalling conditions for the maximum rate of transmission of information with a given proportion of errors.

Synchronous telegraph systems using the 5-unit teleprinter code are described, and a brief account is given of the methods by which the optimum number of signalling conditions may be used for transmission over the cable whilst retaining the advantages of 5-unit working at the terminals.

LIST OF SYMBOLS

ω = Angular frequency.
ω_1 = Cut-off frequency of equalizer.
$Y_1(\omega)$ = Frequency-response characteristic of continuously-loaded cable.
$Y_2(\omega)$ = Frequency-response characteristic of unloaded cable.
$N_1(\omega)$ = Noise power at receiving end of cable as a function of frequency.
$N_2(\omega)$ = Noise power at output of equalizer as a function of frequency.
R = Total resistance of cable.
C = Total capacitance of cable.
L = Total inductance of cable.
I = Information, bits.
t = Time, sec.
n = Telegraph speed, bauds.
S = Number of signalling conditions.
N_0 = Total noise power.
m_1, m_2 = Number of elements in group.

(1) INTRODUCTION

The telegraph circuits with which Great Britain is associated fall into two distinct categories. In the first are those in which the terminal equipment represents a high proportion of the capital cost, such as the inland telegraph circuits and the submarine cable circuits to the Continent. These use the 5-unit teleprinter code in which each character is preceded by a "start" pulse and followed by a "stop" pulse, so that, in effect, the code becomes a 7½-unit one [Fig. 1(*a*)].

With long-distance submarine-cable circuits the capital cost of the terminal equipment represents a much smaller proportion of the whole, so that a greater advantage is taken of the available bandwidth, even if this leads to increased complication of the terminal equipment. With cables owned by British companies, the 3-condition cable code is used almost exclusively, whilst on other cables, notably the American ones, the 5-unit synchronous code is used [Fig. 1(*b*)].

Recent increases in the number of oversea telegraph circuits, together with the extended use of leased-circuit and telex facilities, have led to a need for direct interconnection of internal and oversea circuits. Steps are being taken to achieve this in the case of the long-distance radiotelegraph circuits,[1] and still further advantages would result if submarine-cable circuits could also be connected to the inland networks without the need for manual conversion.

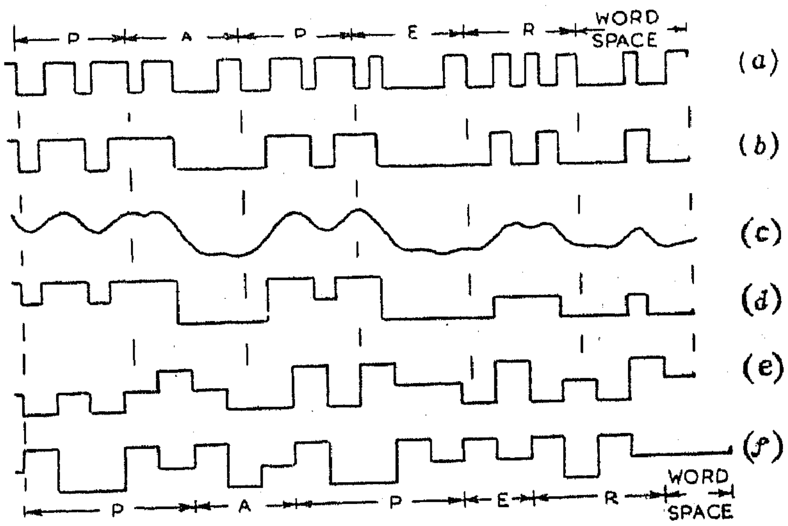


Fig. 1.—Specimen signals showing the different signalling systems and codes.

(*a*) 5-unit start-stop.
(*b*) 5-unit synchronous.
(*c*) 5-unit suppressed singles (received signal).
(*d*) 5-unit, singles suppressed at sending end.
(*e*) Signals from (*b*) converted to 3-condition system.
(*f*) Cable.

(2) CHARACTERISTICS OF SUBMARINE CABLES AND APPARATUS

(2.1) Frequency-Response Characteristic

If leakage is neglected the frequency response of a homogeneous cable terminated in its characteristic impedance is given by

$$Y_1(\omega) = \exp - \left\{\tfrac{1}{2}\omega C\left[(R^2 + \omega^2 L^2)^{1/2} - \omega L\right]\right\}^{1/2} \quad . \quad (1)$$

For an unloaded cable the inductance may be neglected and the response becomes

$$Y_2(\omega) = \exp - (\tfrac{1}{2}\omega CR)^{1/2} \quad . \quad . \quad . \quad . \quad (2)$$

In this case the response depends on the product of the resistance and capacitance, or *CR* value, of the cable.

Fig. 2 shows the frequency-response characteristic of typical submarine cables about 1 000 miles long. Since most cables at present in use are of the unloaded type, the need for efficient use of the available bandwidth is apparent. Continuous loading[2] enables the working bandwidth to be increased to about 100 c/s, and a number of cables of this type were laid in the period 1924–30. No lump-loaded deep-sea cables are known to exist, but recent advances in cable manufacturing techniques and magnetic materials would probably enable lump-loading to be used with advantage in future cables.

(2.2) Apparatus[3, 4]

The terminal apparatus used on submarine cables will not be described in detail except in so far as its characteristics affect the choice of the signalling system. Fig. 3 shows the layout of

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
Mr. Hughes is in the Post Office Engineering Department, Radio Experimental and Development Branch, and was formerly with Cable and Wireless, Ltd.

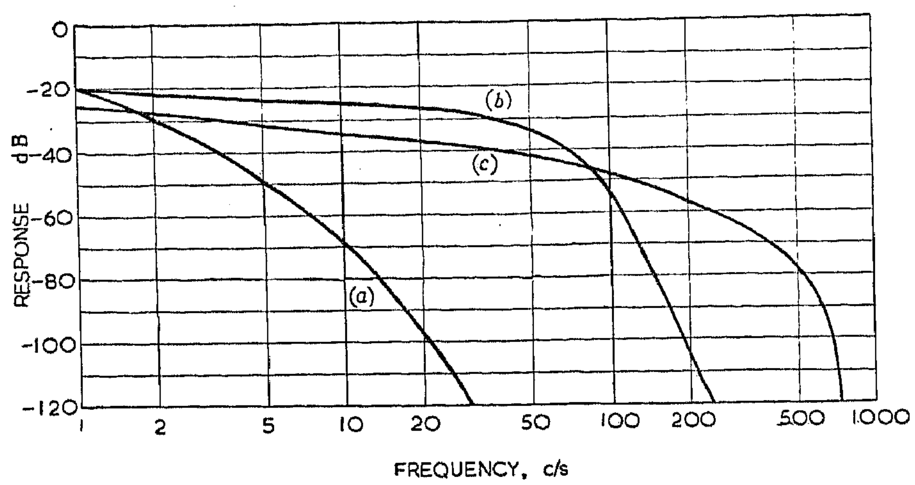


Fig. 2.—Frequency-response characteristics of (*a*) unloaded, (*b*) continuously-loaded and (*c*) lump-loaded submarine cables.
Spacing between loading coils is 2 miles.

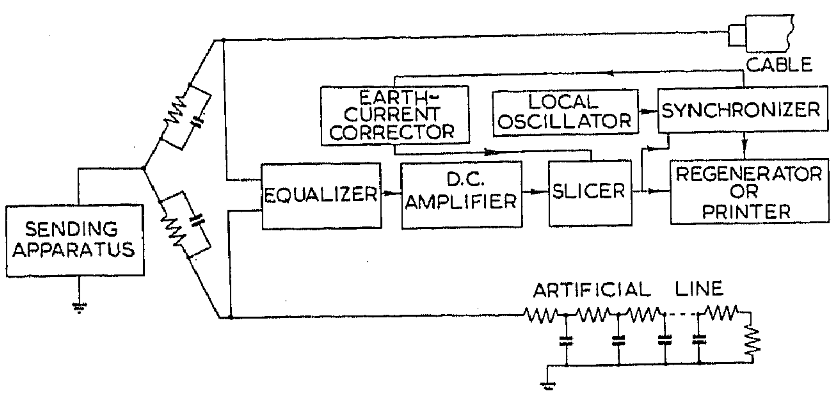


Fig. 3.—Arrangement of apparatus for bridge-duplex operation.

apparatus for a cable worked on the duplex principle, with a bridge network used to prevent the outgoing signals from interfering with the signals from the distant station. For simplex working the bridge network is omitted.

In order to obtain a recognizable telegraph signal it is necessary to equalize the amplitude and phase response of the cable up to a frequency of about 1·6 times the reversal frequency of the signal.[5] An idealized response for an equalizer for an unloaded cable terminated in its characteristic impedance is shown in Fig. 4. The form of a single baud-element pulse after passing

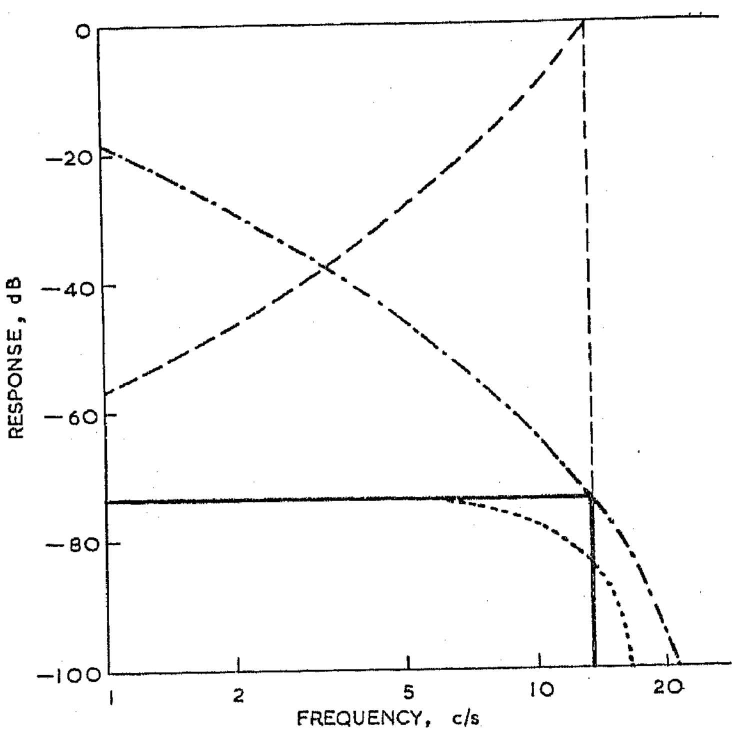


Fig. 4.—Idealized frequency response of equalizer for unloaded cable.
—·—·— Cable response.
— — — Equalizer response.
———— Resultant response of cable and equalizer.
·········· Resultant response modified to avoid ringing.

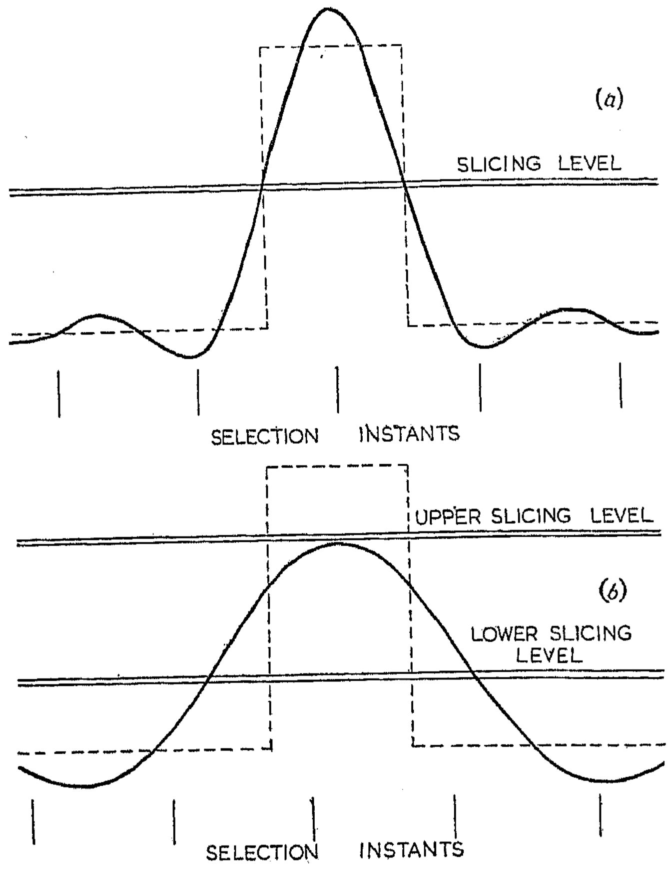


Fig. 5.—Response of cable and equalizer to square baud-element pulse.
———— Received pulse.
– – – – – Transmitted pulse.
(*a*) Cable equalized up to 1·6 times the reversal frequency.
(*b*) Cable equalized up to 0·8 times the reversal frequency.

through the cable and equalizer is shown in Fig. 5(*a*), and from this may be obtained the form of the signal for any combination of elements. In practice, the response of the equalizer is modified to give an overall response such as that shown by the dotted line in Fig. 4. This is preferable, since the ringing produced by a sharp cut-off would cause interference with succeeding baud-element pulses.

From the amplifier the signal passes to the slicing stages and thence to the selecting apparatus. The incoming signals are themselves used to keep the selecting apparatus in the correct phase relationship to the signals by means of the synchronizer.

(2.3) Earth Currents

The differences in potential which exist between different points on the earth's surface cause earth currents to flow along the cable. One method of compensating for the effect of earth currents is to use the information contained in the phase difference between the local oscillator and the incoming signals, together with the polarity of the signal, to alter the slicing level when necessary.[3] The effect of earth currents is also eliminated if a.c. coupling is used between the cable and the amplifier. In this case a form of d.c. restoration or "local correction"[4] is used to restore the low-frequency components of the signal.

(2.4) Noise

Where simplex working is employed, and in some cases of duplex working, noise arises from atmospheric disturbances, power sources and adjacent cables. For disturbances transmitted through the ocean the higher frequencies are attenuated much more rapidly than the lower frequencies.[6,7,8] The result is that, with the possible exception of peaks at power frequencies, the greater part of the noise energy present at the input to the equalizer is concentrated in the low-frequency part of the spectrum. When an offshore submerged repeater[8] is inserted

in the cable the effect is even more marked, since, apart from valve noise, the disturbances from the atmosphere have to penetrate a considerable depth before reaching the cable.

It is difficult to estimate the noise spectrum on theoretical grounds, but a very rough approximation may be made by supposing that a noise generator, whose output is uniform over the range of frequencies under consideration, is located at a point some distance along the cable. Then the noise power spectrum at the receiving end of an unloaded cable terminated in its characteristic impedance may be expressed in the form

$$N_1(\omega) = \exp\left[-2k\sqrt{(\tfrac{1}{2}\omega CR)}\right] \quad . \quad . \quad . \quad (3)$$

where k is a factor which is determined by the distance of the supposed noise source from the receiving end.

The main source of noise where duplex working is used is usually the interference from the outgoing signals, caused by the lack of balance in the bridge. Although it is possible to obtain an almost perfect balance at one or more frequencies, there is inevitably some lack of balance at intermediate frequencies. The degree of unbalance varies with frequency in an irregular manner, and Fig. 6 shows a typical result. The spectrum of a random

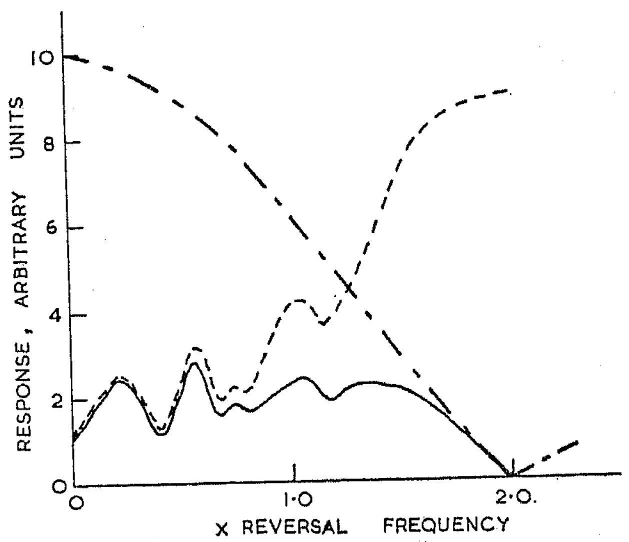


Fig. 6.—Typical variation of duplex out-of-balance with frequency.
— — — Duplex out-of-balance.
—·—·— Spectrum of random telegraph signal.
———— Resultant noise spectrum.

telegraph signal[5,9] is also plotted, so that the resultant noise spectrum follows the full line in Fig. 6. This approximates roughly to a uniform noise spectrum at the input to the equalizer.

(3) SIGNALLING CONDITIONS AND CODES

It is convenient to consider first the number of signalling conditions and the code which should be used in order to transmit information at the maximum rate and with an acceptable proportion of error. This ideal cannot always be realized on a practical system, but it does provide a useful measure against which the practical system may be compared.

(3.1) Number of Signalling Conditions

The rate of transmission of information is given by

$$\frac{I}{t} = n \log_2 S \text{ bits per second} \quad . \quad . \quad . \quad (4)$$

The two factors n and S are mutually dependent, since an increase in the number of signalling conditions usually involves a reduction in the baud speed to ensure the same degree of protection against noise. Whether an overall improvement will result from an increase in the number of signalling conditions depends on the noise spectrum and the response characteristic of the cable.

If a noise spectrum such as that given by eqn. (3) is present at the input to the unloaded-cable equalizer, the spectrum at the output of the equalizer will have the form

$$N_2(\omega) = \exp\left[(1-k)\sqrt{(2\omega CR)} - \sqrt{(2\omega_1 CR)}\right] \quad . \quad . \quad (5)$$

The total noise power at the output of the equalizer will then be given by

$$N_0 = \exp\left[-\sqrt{(2\omega_1 CR)}\right]\int_0^{\omega_1} \exp\left[(1-k)\sqrt{(2\omega CR)}\right]d\omega \quad (6)$$

If the baud speed for an acceptable error rate with a given number of signalling conditions is known, it is then possible to determine whether an increase in the number of signalling conditions will give an increase in information-carrying capacity with the same error rate.

(3.1.1) Duplex Operation.

Figs. 7 and 8 show how the signal/noise ratio at the equalizer output varies with baud speed for an unloaded cable

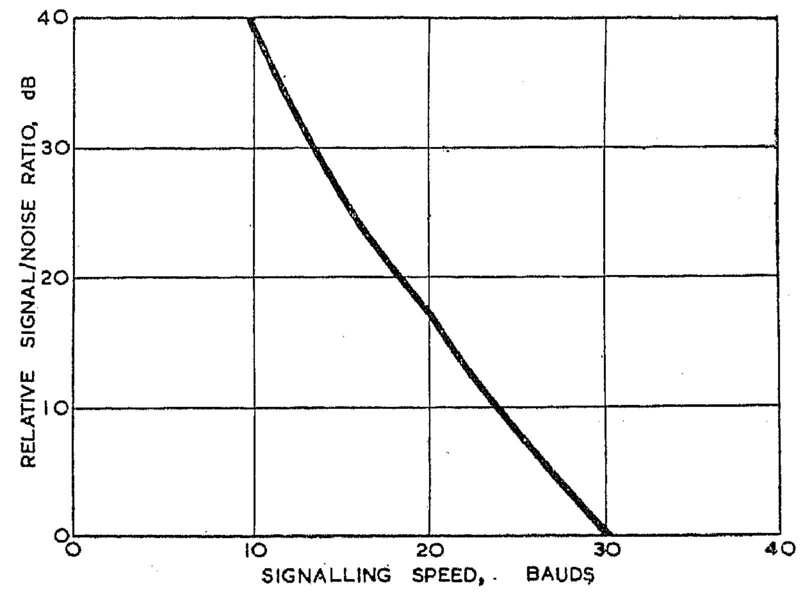


Fig. 7.—Variation of relative signal/noise ratio with signalling speed for assumed noise spectra. (Cable worked duplex.)

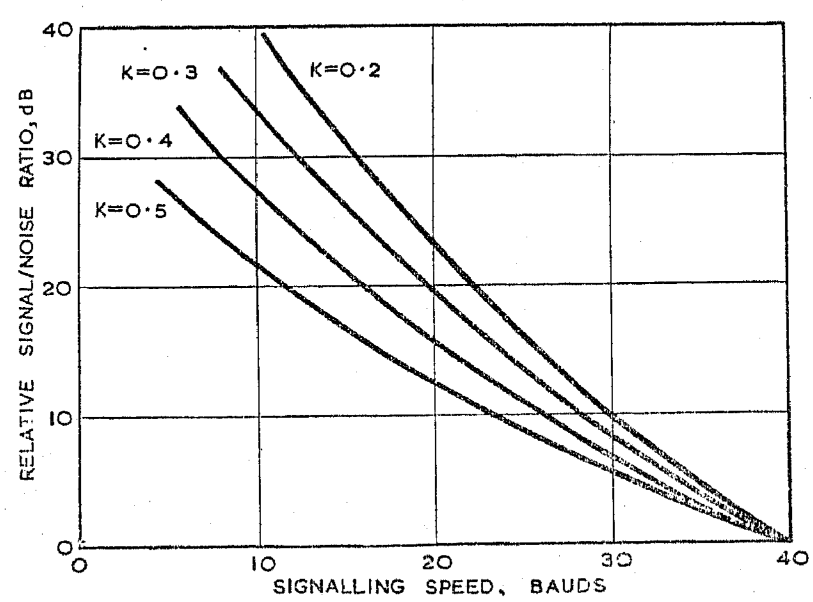


Fig. 8.—Variation of relative signal/noise ratio with signalling speed for assumed noise spectra. (Cable worked simplex.)

($CR = 2\cdot 0$ sec) terminated in its characteristic impedance and equalized up to a frequency of $1\cdot 6$ times the reversal frequency. For such a cable, under duplex conditions (Fig. 7), a working speed of 30 bauds might be achieved with two signalling conditions. If the number of signalling conditions were increased to three, and the speed reduced to 26 bauds to give the same signal/noise ratio between adjacent signalling conditions, the rate

of transmission of information could be increased by about 38%. With four signalling conditions the theoretical gain is approximately 57%. The increase in the number of signalling conditions increases the practical difficulties in compensating for earth currents and, in some cases, for variations in leakance faults on the cable.

(3.1.2) **Simplex Operation.**

For simplex circuits the total noise power at the output of the equalizer may be found approximately by substituting the appropriate value of k in eqn. (6). It will be seen from Fig. 8 that, for an estimated working speed of 40 bauds with two signalling conditions, there would be no advantage in increasing the number of signalling conditions for values of k greater than about $0 \cdot 5$. An exception may arise in some cables where it would be advantageous to increase the number of signalling conditions in order that the equalizer may be designed to cut off below power frequencies.

(3.1.3) **Loaded Cables.**

As the signalling frequency is increased on a loaded cable there comes a point at which the fall in the response becomes much more rapid, as shown in Fig. 2. Once this frequency is reached the only practical way of materially increasing the information-carrying capacity of the cable is to increase the number of signalling conditions.

(3.2) Codes

Once the optimum number of signalling conditions has been determined, there remains the problem of coding the information from the telegraph message source to take advantage of the available capacity. A minimum of 37 characters is necessary for the transmission of English words and numerals, and if the other symbols commonly used in telegrams are included, this figure is increased to 45. From sample batches of telegrams transmitted on oversea circuits, values for the relative frequency of occurrence of the characters were obtained. With Huffman's method[10] for producing a minimum-redundancy code it was found that, for the telegraph message source, an average of $4 \cdot 7$ elements per character would be required on a 2-condition signalling system. If a 3-condition system were used, an average of $2 \cdot 95$ elements per character would be required.

It is doubtful whether the complex code convertors and information storage systems required to convert from a 5-unit to a Huffman code would be justified in practice, since, even if allowance is made for case shifts in the 5-unit code, the gain in capacity is only 15% for a 2-condition system.

(4) FIVE-UNIT SYNCHRONOUS SYSTEMS

A number of systems are in use in which the 5-unit teleprinter code is used for signalling on the cable. Owing to the need for economy of bandwidth, the "start" and "stop" pulses are omitted and the signals themselves are used to keep the selecting mechanism at the receiving end in the correct phase relationship.

One disadvantage with systems which do not transmit the "start" and "stop" pulses is that supervisory signals, which are normally indicated by the continuous "mark" or continuous "space" conditions, cannot be transmitted.

(4.1) Five-Unit Fully-Formed Systems

The simplest form of 5-unit synchronous system is one in which the equalizer network is so designed that frequencies up to about $1 \cdot 6$ times the reversal frequency of the signal are passed whilst the higher frequencies suffer considerable attenuation. Then, provided that the phase equalization is correct, the signal elements may be selected by the receiving apparatus without difficulty, whilst the bandwidth is such that no unnecessary noise is admitted.

(4.2) Five-Unit Suppressed-Singles Systems[11]

If the maximum frequency of response of the system is reduced to about $0 \cdot 8$ times the reversal frequency, it is possible to reinsert artificially the higher signalling frequencies at the receiving end provided that a third signalling condition, midway between the "mark" and "space" conditions, is recognized by the receiving apparatus. When such a condition is recognized it signifies that the element to be inserted is one of opposite sign to that last received. The received signals take the form shown in Fig. 1(*c*).

The effect of the reduction in bandwidth on a single baud-element pulse is shown in Fig. 5(*b*). If the phase of the selecting instants were such that a selection took place at the centre of the pulse, the selection for the preceding element would occur as the waveform crossed the lower slicing level, and a wrong selection would be probable. If is therefore necessary to advance the phase of the selecting instants as shown in the Figure.

If only a single baud-element pulse is considered it is found that, at the selection instant, the signal amplitude relative to the slicing level is some 10 dB below that of the fully-formed signal. However, a comparison of Figs. 5(*a*) and 5(*b*) shows that in the suppressed-singles system a greater disturbance from the residue of preceding baud-element pulses may be expected. By consideration of the effect of various combinations of preceding pulses it is found that the effective signal amplitude is approximately 15 dB below that of the fully-formed signal.

Although the disturbance from the preceding baud-element pulses may be reduced by making the equalizer cut-off less sharp, as explained in Section 2.2, this also has the effect of increasing the rise time of the pulse. For the fully-formed signals, where the selections take place at full amplitude, a slight increase in rise time has little effect. However, in the suppressed-singles system some selections take place during the rise of the pulse, so that an increase in the rise time reduces still further the effective signal amplitude at the selection instant. The result is that, even for the modified equalizer response, the effective signal amplitude is still about 15 dB below that of the fully-formed signal.

With reference to Fig. 7, for the unloaded duplex cable with a uniform noise spectrum the speed would have to be reduced from 30 bauds to about 21 bauds to give the same protection against noise as in the fully-formed case. This represents the speed of the received signals with the singles suppressed, so that the speed of the transmitted signals is about 42 bauds. The net gain under these conditions is thus approximately equal to the gain obtained with a full 3-condition system. For simplex working the 3-condition system shows a greater gain.

A disadvantage of this system is that, if a reduction in speed should be required, it is necessary to alter the characteristic of the equalizer network to prevent the reversal frequency from being passed to the selection mechanism at sufficient amplitude to cause a mis-selection.

(4.3) Single Elements suppressed at Sending End

Instead of the single elements being suppressed by the rejection of the higher frequencies in the cable and equalizer, it is possible to suppress them by means of the signalling apparatus at the sending end. The transmitted signal is then as shown in Fig. 1(*d*). The advantage of this system is that the circuit speed may be reduced without altering the characteristic of the equalizer network.

The amplitude of single elements is reduced still further, since these are transmitted as half-amplitude pulses. The result is that the gain in capacity is somewhat less than when single elements are suppressed by the cable and receiving network.

(5) CODE CONVERTORS FOR INCREASING THE NUMBER OF SIGNALLING CONDITIONS

When the increase in the number of signalling conditions gives an increase in information carrying capacity, code convertors are necessary to enable connection to the teleprinter system to be made.

(5.1) Five-Unit to Cable-Code Convertors

Convertors suitable for converting from 5-unit code the 3-condition cable code and vice versa have been described by Carter and Wheeler.[12] Owing to the unequal length of characters in the cable code, some form of storage is necessary, and this usually takes the form of perforated tape. Although tape is used at the sending end for nearly all synchronous systems, the stage at the receiving end constitutes an additional delay in message handling.

For an unequal-length code the average length of a cable-code character (3·7 elements) does not compare very favourably with the ideal 3-condition Huffman code (2·95 elements per character). The redundancy of the cable code makes it, to a certain extent, error-detecting and also it is readable on pen records.

(5.2) Equal Element-Group Code Convertors

It is possible to convert from a 2-condition code to one employing a greater number of signalling conditions by converting a fixed number of elements, not necessarily a complete character, of the 2-condition code into a smaller number of elements of the new code. Since the information is proportional to the logarithm of the number of signalling conditions, the ideal convertor should be designed to convert groups of elements in the ratio

$$\frac{m_2}{m_1} = \frac{\log S}{\log 2} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

where m_2 elements of the 2-condition code are converted to m_1 elements of the code employing S signalling conditions.

It is not always possible to achieve this ideal ratio in practice. One system for use on cable circuits employing three signalling conditions is designed to convert a group of three elements of the 2-condition (5-unit) code to a group of two elements of a 3-condition code, as shown in Table 1. This falls short of the ideal ratio by only 6%, and even then the spare combination may be used for the transmission of supervisory signals.

Table 1

CONVERSION OF 2-CONDITION (5-UNIT) CODE TO 3-CONDITION SYSTEM

Two-condition group	Three-condition group
MMM	0+
SSS	0−
MMS	++
MSM	+−
MSS	+0
SMS	−+
SSM	−−
SMM	−0
Supervisory signal	00

The code conversion may be performed by electronic,[14] mechanical or electro-mechanical[13, 15] means. In the electronic convertor the elements are stored on cold-cathode trigger valves, and the switching operations required to perform the conversion are carried out by a system of selenium diodes. In an electro-mechanical system polarized relays are used for the element storage, and the switching is carried out by the relay contacts.

(6) CONCLUSIONS

The 5-unit teleprinter code is so well established that submarine cable circuits should be designed to accept and deliver this code at the terminals. This does not rule out the use of code convertors if an appreciable gain in information-carrying capacity is thereby obtained.

For loaded cables and duplex unloaded systems both the 3-condition and receiving-end suppressed-singles methods have been proved to be superior to the fully-formed 2-condition method. The information-carrying capacities of the 3-condition and suppressed-singles methods are almost equal, and prolonged trials under working condition would be necessary to determine which is the better. The 3-condition system appears to have an advantage in that it provides a better signal for synchronization and earth-current correction; moreover it facilitates the transmission of supervisory signals.

So far as is known, methods using a greater number of signalling conditions have not yet been tried, and again, tests on working circuits would be necessary to ascertain whether the increase in information-carrying capacity would be sufficient to compensate for the reduced stability in working.

For unloaded cables with simplex working there would appear to be little advantage in increasing the number of signalling conditions above two, except for the purpose of standardization.

(7) ACKNOWLEDGMENTS

The author wishes to express thanks to the Engineer-in-Chief, Cable and Wireless, Ltd., for permission to publish the paper, and to Mr. R. L. Saunders for his advice during its preparation.

(8) REFERENCES

(1) HAYTON, T., HUGHES, C. J., and SAUNDERS, R. L.: "Telegraph Codes and Code Convertors," *Proceedings I.E.E.*, Paper No. 1585 R, November, 1953 (**101**, Part III, p. 137).

(2) FOSTER, A. E., LEDGER, P. G., and ROSEN, A.: "The Continuously-Loaded Submarine Cable," *Journal I.E.E.*, 1929, **67**, p. 475.

(3) WOOD, K. L.: "Empire Telegraph Communications," *ibid.*, 1939, **84**, p. 638.

(4) CRAMER, C. H.: "Some Modern Techniques in Modern Cable Telegraphy," *Transactions of the American I.E.E.*, 1947, **66**, p. 765.

(5) MILNOR, J. W.: "Submarine Cable Telegraph," *ibid.*, 1922, **41**, p. 20.

(6) MILNOR, J. W.: "Submarine Cable Telegraphy; Influence of Interference," *ibid.*, 1931, **50**, p. 397.

(7) GILBERT, J. J.: "Extraneous Interference on Submarine Telegraph Cables," *Bell System Technical Journal*, 1926, **5**, p. 404.

(8) CRAMER, C. H.: "Submerged Repeater for Ocean Cable," *Electrical Engineering*, 1951, **70**, p. 566.

(9) BELL, D. A.: "Statistical Methods in Electrical Engineering" (Chapman and Hall, London, 1953).

(10) HUFFMAN, D. A.: "A Method for the Construction of Minimum Redundancy Codes," *Proceedings of the Institute of Radio Engineers*, 1952, **40**, p. 1098.

(11) ANGEL, H.: "Printing Telegraphs on Non-Loaded Ocean Cables," *Transactions of the American I.E.E.*, 1927, **46**, p. 884.

(12) CARTER, R. O., and WHEELER, L. K.: "Code Convertors for the Interconnection of Morse and Teleprinter Systems," *Proceedings I.E.E.*, Paper No. 1570 R, October, 1953 (**101**, Part III, p. 151).

(13) NELSON, C. F.: U.S. Patent No. 2046964, 1936.

(14) HUGHES, C. J.: British Patent Application No. 25682, 1954.

(15) SAUNDERS, R. L., and HUGHES, C. J.: British Patent Application No. 437, 1955.

Paper No. 1921 R
Nov. 1955

621.317.328.026.6

A NEW PERTURBATION METHOD FOR MEASURING MICROWAVE FIELDS IN FREE SPACE

By Professor A. L. CULLEN, Ph.D., Associate Member, and J. C. PARR, B.Sc.

(*The paper was first received* 26*th May, and in revised form* 15*th July,* 1955).

SUMMARY

A perturbation method for measuring free-space microwave fields is described. A short thin metal rod forms the perturbing element. In order to avoid confusing the reflection from this rod with unwanted reflection the rod is arranged to spin about an axis perpendicular to its length. The perturbation of the field at the source is therefore modulated at a characteristic frequency, and by using an a.c. amplifier, steady reflections can be eliminated. The apparatus described operates at a wavelength of about 3·2 cm. Experimental results are given which verify the theory for linearly polarized waves. Extension of the method to elliptically polarized waves is possible in principle, but inherent ambiguities may be troublesome in practice.

LIST OF SYMBOLS

A, B = Forward and backward travelling-wave amplitudes in waveguide.
a, b = Waveguide dimensions.
E_1, E_2 = Electric fields incident on, and reflected by, the scattering rod.
F = A complex vector function of position such that $E_1 = AF$.
G = Amplifier gain.
H_1, H_2 = Magnetic fields associated with E_1, E_2.
I = Rectified crystal current.
i, j, k = Unit vectors along x, y, and z axes, respectively.
$j = \sqrt{-1}$.
J = Current density.
K = A constant of proportionality.
k = Phase constant $= 2\pi/\lambda$.
l = Length of short dipole.
M = Dipole moment.
n = Unit normal.
q = Electric charge.
R_1, R_2 = Resistances.
r = Radial distance.
u = Unit vector.
w = Vector function of position, characteristic of waveguide mode used.
x, y, z = Cartesian co-ordinates.
Y_0, Z_0 = Characteristic admittance and impedance of H-mode in a waveguide.
α = Polarizability of dipole.
ϵ_0 = Permittivity of free space.
θ = Orientation of electric field.
λ = Wavelength.
μ_0 = Permeability of free space.
ρ = Reflection coefficient.
ϕ = Phase-shifter setting.
ψ = Brush setting on commutator.
Ω = Angular velocity of scattering rod.
ω = Radian frequency of the electromagnetic waves.

(1) INTRODUCTION

Perturbation methods have been proposed and used for measuring microwave fields in cavities by several workers,[1-4] but to the authors' knowledge only once has this principle been applied to the measurement of radiation fields. In these measurements[3] a small scattering element was introduced into the mouth of a horn in order to determine the aperture-field distribution under transmitting and receiving conditions.

To measure the field strength at any point in a transmitting horn by this method the unperturbed horn is initially matched, and the input standing-wave ratio is then measured with the scattering element in position. The associated reflection coefficient is related to the unperturbed field strength in the region occupied by the scattering element, and the precise relationship can be obtained by applying the reciprocity theorem to the system.

In principle, this method could be used to measure the electromagnetic field distribution at any distance from a transmitting horn; in practice, however, the returning wave rapidly becomes weaker as the distance increases, and its detection is difficult. Moreover, it is not easy to separate this scattered wave from waves scattered by other neighbouring objects, possibly including the observer.

In spite of these difficulties, a perturbation method of field measurement is more attractive than a direct method, because the disturbance of the field to be measured can usually be made smaller. This is partly due to the fact that the scattering element is not electrically connected to the remainder of the apparatus, whereas in the direct method there are d.c. or r.f. leads to the apparatus which must further disturb the field being measured, and partly because the scattering element itself can be made much smaller in volume than the usual type of crystal detector unit.

For these reasons a new perturbation method has been developed which, because of the small size of the scattering element, is particularly suitable for measuring the "fine detail" of an electromagnetic field, e.g. the field near the focal point of a microwave lens or in a diffracting aperture. It should have numerous similar applications.

(2) A SIMPLE PERTURBATION METHOD

Consider the arrangement of apparatus shown in Fig. 1. One of the two symmetrical arms of a matched hybrid-T waveguide junction is terminated by a perfectly-matched load, the other by a horn which is perfectly matched to free space. The shunt arm of the hybrid T is fed from a reflex-klystron oscillator, and the series arm terminates in a crystal detector. To simplify the discussion it will be assumed for the present that the oscillator and crystal are also perfectly matched to the waveguide.

Because of the symmetry of the hybrid T, the crystal detector is completely decoupled from the oscillator in the absence of a scattering element. However, when the latter is introduced into the field a wave is scattered back into the horn. This wave travels back along the waveguide from the horn to the hybrid T where it shares its power between the crystal detector and the

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.

Prof. Cullen and Mr. Parr were formerly in the Department of Electrical Engineering, University College, London.
Prof. Cullen is now Professor of Electrical Engineering, University of Sheffield.
Mr. Parr is now with Kelvin Hughes, Ltd.

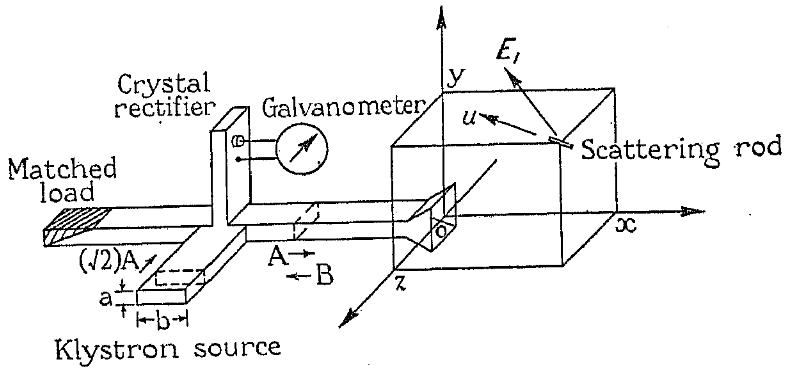


Fig. 1.—Arrangement of apparatus for simple perturbation method.

source. The rectified current flowing in the crystal circuit is a measure of field strength at the scattering element.

A precise quantitative analysis of the situation can be made using the Lorentz reciprocity theorem; the details of the calculation are given in Section 9. The electric field alone is assumed to be effective in exciting the scattering element; this is the case of interest in later work. Consider an outgoing wave whose complex amplitude at the centre of the rectangular waveguide in the reference plane shown in Fig. 1 is denoted by A. This wave will produce an electric field $E_1(x, y, z)$ in the free-space region outside the horn.

Since the amplitude of the field $\boldsymbol{E}_1$ is directly proportional to A, it is convenient to separate the amplitude and space-variation of $\boldsymbol{E}_1$ by writing

$$\boldsymbol{E}_1(x, y, z) = A\boldsymbol{F}(x, y, z) \quad . \quad . \quad . \quad . \quad (1)$$

when the quantity to be determined is now the function $\boldsymbol{F}$ rather than $\boldsymbol{E}_1$. It should be noted that $\boldsymbol{F}$ is a dimensionless complex vector function of position, since both A and $\boldsymbol{E}$ represent electric field strengths.

Now suppose that a thin metal rod of electric polarizability* α and of zero magnetic polarizability is placed with its centre at the point (x, y, z) and with its axis in the direction of the unit vector $\boldsymbol{u}$.

Some of the incident radiation will now be scattered back into the horn, and this will give rise to a backward-travelling wave whose complex amplitude in the reference plane of Fig. 1 we shall denote by B. An application of the Lorentz form of the reciprocity theorem, carried out in Section 9.1, shows that

$$B = A(j\omega\alpha Z_0/ab)(\boldsymbol{u} \cdot \boldsymbol{F})^2 \quad . \quad . \quad . \quad . \quad . \quad (2)$$

If the crystal has a square-law relationship between rectified current and r.f. field, the rectified current I is proportional to $|B|^2$, or

$$I = K|B|^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

Combining eqns. (2) and (3) we see that the rectified current is given by

$$I = K\left|\frac{\omega\alpha Z_0}{ab}\right|^2 |\boldsymbol{u} \cdot \boldsymbol{F}|^4 |A|^2 \quad . \quad . \quad . \quad . \quad (4)$$

Thus if the rectified current is measured with the centre of the rod at any specified point, and with the axis of the rod successively parallel to the x, y and z axes (i.e. $\boldsymbol{u} = \boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$ in turn), the relative magnitudes of the complex cartesian components F_x, F_y and F_z of the vector function can be found. If this process is repeated for other points the relative magnitudes of the components of electric field strength can be found as a function of position.

The following points relevant to a practical application of this scheme should be noted. First, the scattering rod must be supported in the field in some way, e.g. by a thin nylon cord. This cord will probably have negligible scattering effect, but the supports to which it is attached may produce an appreciable amount of scattered radiation. This, and any other unwanted scattered radiation, will affect the measurements in two distinct ways:

(a) It will contribute directly to the field scattered back into the horn.

(b) It will modify the field at the point of observation.

Secondly, the rectified current will usually fall off extremely rapidly with increasing distance from the source. Thus, for a simple horn of the kind shown in Fig. 1, the field will fall off as $1/r$ with increasing radial distance r from the centre of the aperture of the horn, when r is sufficiently large in comparison with the aperture dimensions. Thus the function F in eqn. (4) is inversely proportional to the distance, and the rectified current is inversely proportional to the fourth power of the distance. Finally, the system gives no indication of the phases of the field components.

In Section 3 an improved form of the apparatus will be described in which, with the exception of (b) above, all these defects are remedied.

(3) A SPINNING-DIPOLE SYSTEM FOR LINEARLY POLARIZED WAVES

To distinguish between the signal reflected from the scattering dipole and other unwanted reflections, the dipole can be rotated at a constant speed about an axis perpendicular to its length so that the reflected waves fluctuate in a characteristic manner.

In Fig. 2, suppose that the dipole is always parallel to the

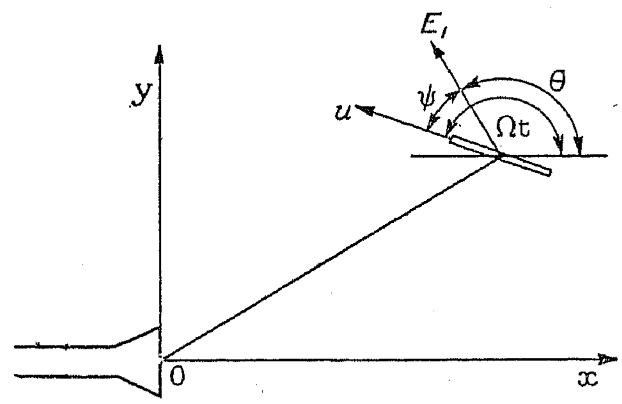


Fig. 2.—Diagram illustrating notation for spinning-dipole method.

xy-plane, and spins about an axis parallel to the z-axis with angular velocity Ω. If we restrict attention for the present to linearly polarized waves, the product $\boldsymbol{u} \cdot \boldsymbol{F}$ in eqn. (2) can be written $F_t \cos\psi = |F_t|\varepsilon^{j\phi_t}\cos\psi$ where $\psi = \Omega t - \theta$, as can be seen from Fig. 2, and $\boldsymbol{F}_t(x, y, z)$ is the transverse part of $\boldsymbol{F}$, i.e. the projection of $\boldsymbol{F}$ on the xy-plane. The assumption of linear polarization means that the x and y components of $\boldsymbol{F}$ have the same phase, so that a single phase angle ϕ_t is sufficient.

Eqn. (2) for B then becomes

$$B = j\frac{\omega\alpha Z_0}{ab}|F_t|^2\varepsilon^{j2\phi_t}A\cos^2(\Omega t - \theta) \quad . \quad . \quad (5)$$

The instantaneous value of the rectified current, i, is found by substituting eqn. (5) in eqn. (3).

From trigonometrical formulae the current can be expressed as the sum of d.c., double-frequency and quadruple-frequency components; by means of a tuned a.c. amplifier or suitable filter the d.c. and quadruple-frequency components can be rejected,

* The polarizability α is a real number if the rod is short in comparison with the wavelength, and quasi-static treatment is adequate. For longer rods there will be a phase difference between the applied electric field and the induced dipole moment. The concept of polarizability can be usefully retained, however, if we allow α to be a complex number to account for this phase difference.

and the double-frequency component alone is then significant. Its value is found to be

$$I_2 = K\left|\frac{\omega\alpha Z_0}{ab}\right|^2 |F_t|^4 |A|^2 \tfrac{1}{2}\cos 2(\Omega t - \theta) \quad . \quad . \quad (6)$$

If the current I_2 given in eqn. (6) is compared in phase with a low-frequency, reference current derived from the shaft driving the spinning dipole and varying with time as $\cos 2\Omega t$, the phase difference θ can be found, and hence the direction of the electric field in space is determined. If the field is elliptically polarized a more complicated situation exists, and this is considered in more detail in Section 6.

In the derivation of eqn. (6), it has been assumed that unwanted reflections are negligible. It is clear that such reflections, although they cannot contribute directly to the double-frequency part of the reflected waves, may still modify the corresponding rectified current, because of the square-law characteristic of the crystal.

This difficulty can be overcome by adding an r.f. reference current of constant and relatively large magnitude to the existing r.f. current in the crystal by deliberately unbalancing the hybrid T. Any residual unwanted reflections will merely add to this reference signal, and will usually be negligible in comparison with it. A reference signal can be provided by putting a mismatching screw in the arm previously terminated in a matched load. At the same time a calibrated reflectionless phase-shifter is introduced,

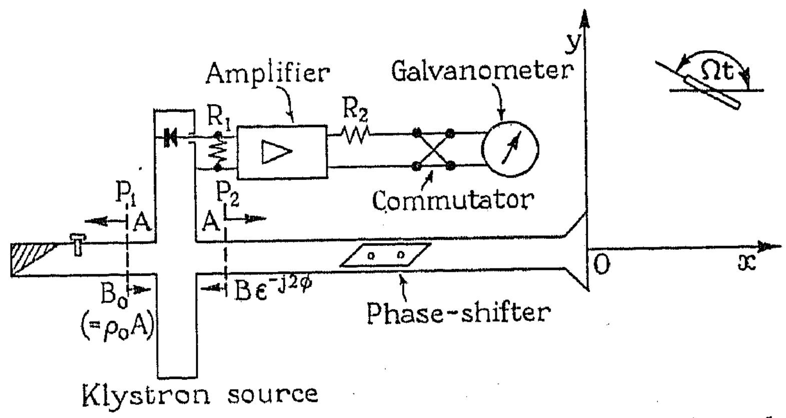


Fig. 3.—Arrangement of apparatus for spinning-dipole method.

as shown in Fig. 3. If the phase lag due to the phase-shifter is ϕ for one-way transmission the returning wave is modified by a phase factor $\varepsilon^{-j2\phi}$ and becomes $B\varepsilon^{-j2\phi}$. With these modifications, the rectified current* is now given by:

$$i = K|B_0 - B\varepsilon^{-j2\phi}|^2 \quad . \quad . \quad . \quad . \quad . \quad (7)$$

in place of eqn. (3), with B given by eqn. (5) as before. The negative sign before B appears because the crystal is in the series arm of the hybrid T. Before proceeding further with the analysis it is instructive to interpret eqn. (7) by means of a time-vector diagram, as shown in Fig. 4. If, as before, the field to be measured is linearly polarized, then as the dipole rotates the vector $-B\varepsilon^{-j2\phi}$ shrinks to zero and then expands to its original size (without changing its direction or its sense) twice per revolution of the dipole. The associated change in the length of the resultant vector is greatest when B and $-B\varepsilon^{-j2\phi}$ are in line, and this is the condition for the double-frequency component of the rectified current to have a maximum value.

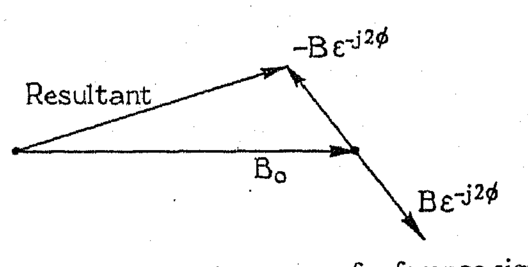


Fig. 4.—Vector diagram illustrating sums of reference signal and signal scattered from dipole.

To confirm this argument analytically is cumbersome for the general case, but if we make use of the fact that $|B|$ is much less than $|B_0|$ in practice, a relatively simple expression for the double-frequency current can be obtained. We therefore rearrange and approximate in eqn. (7) as follows:

$$\begin{aligned} i &= K|B_0|^2\left|1 - \frac{B\varepsilon^{-j2\phi}}{B_0}\right|^2 \\ &\simeq K|B_0|^2\left|1 - \frac{2B\varepsilon^{-j2\phi}}{B_0}\right| \\ &\simeq K|B_0|^2\left[1 - \mathscr{R}\left(\frac{2B\varepsilon^{-j2\phi}}{B_0}\right)\right] \quad . \quad . \quad . \quad (8) \end{aligned}$$

In absence of the scattering element $B = 0$, and so the factor $K|B_0|^2$ in eqn. (8) can be identified with the steady rectified current I_0 which flows in the crystal circuit due to the r.f. reference current alone.

It is also convenient to express B_0 in terms of the reflection coefficient ρ_0 due to the mismatching screw, as shown in Fig. 3. Eqn. (8) then becomes

$$i = I_0\left[1 - \mathscr{R}\left(\frac{2B\varepsilon^{-j2\phi}}{\rho_0 A}\right)\right] \quad . \quad . \quad . \quad . \quad (9)$$

If we substitute for B from eqn. (5) in eqn. (9) the double-frequency component of i becomes:

$$I_2 = -I_0\mathscr{R}\left[\frac{(j\omega\alpha Z_0 \exp 2j(\phi_t - \phi)}{\rho_0 ab}\right]|F_t|^2 \cos 2(\Omega t - \theta) \quad (10)$$

The phase angle of ρ_0 depends on the position chosen for the reference plane P_1 in Fig. 3. It is convenient to choose P_1 so that $\arg(\rho_0) = \arg(j\alpha) + \pi$. The position of P_2 then follows by symmetry. On taking the real part of eqn. (10) we obtain

$$I_2 = I_0\frac{\omega|\alpha|Z_0}{ab|\rho_0|}|F_t|^2 \cos 2(\phi_t - \phi)\cos 2(\Omega t - \theta) \quad . \quad (11)$$

A voltage proportional to I_2 is available at the output end of the amplifier shown in Fig. 3, and is rectified, the rectified current being measured by a galvanometer. Since we also wish to know the phase angle θ it is convenient to make use of a phase-sensitive rectifier, which consists of a mechanical commutator driven by the motor which drives the scattering rod. This is indicated schematically in Fig. 3.

The effect of the commutator is equivalent to multiplying the amplifier output by a "commutating waveform" which takes the form of a square wave of unit amplitude. The "phase" of this square wave can be controlled by changing the position of the brushes on the commutator. The "phase angle" 2ψ, measured from the centre (in time) of a positive half-cycle of this square wave, is indicated in Fig. 5, and is a measure of the brush position. Fig. 5 also shows I_2 as a function of $2\Omega t$. The d.c. component of the rectified output of the phase-sensitive detector is proportional to the average value of the product of the waveforms shown in Fig. 5(*a*) and 5(*b*), the constant of proportionality, $2GR_1/\pi R_2$, containing the gain of the amplifier and the resistances R_1 and R_2 shown in Fig. 3. The result is

$$I_{av} = I_0\left[\frac{2GR_1}{\pi R_2}\left(\frac{\omega Z_0|\alpha|}{ab|\rho_0|}\right)\right]|F_t|^2 \cos 2(\phi_t - \phi)\cos 2(\theta - \psi) \quad . \quad . \quad . \quad . \quad (12)$$

here ψ is a measure of the brush position.

* It should be pointed out at this stage that the assumption of a square law for the crystal characteristic is not essential to the method; any arbitrary functional relationship $i = f|B - B\varepsilon^{-j2\phi}|$ may now be permitted, as an application of Taylor's theorem will show, provided that $|B| \ll |B_0|$.

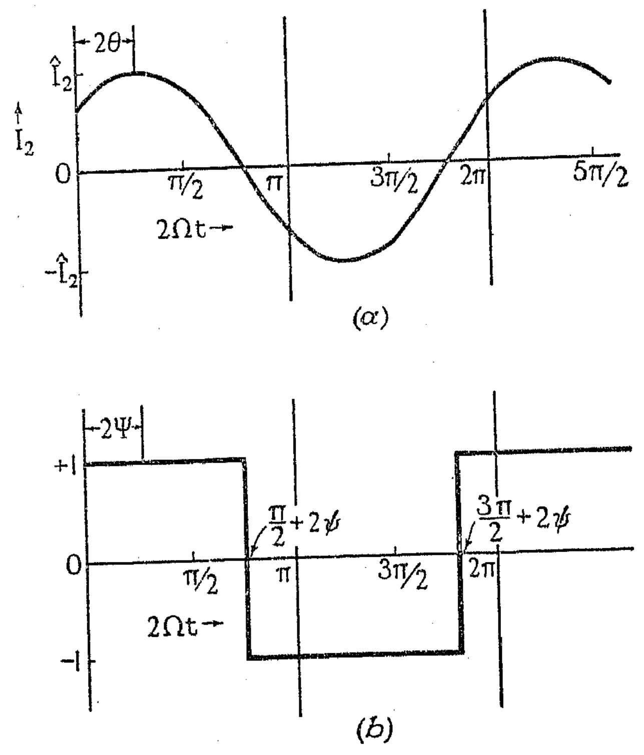


Fig. 5.—Illustrating action of phase-sensitive detection.
(a) Alternating component of rectified current.
(b) Multiplying factor due to commutator.

For a given scattering rod and mismatching screw the factor in square brackets is constant, so that the final form for the average current through the galvanometer can be written

$$I_{av} = I_0 G_1 |F_t|^2 \cos 2(\phi_t - \phi) \cos 2(\theta - \psi) \quad . \quad . \quad (13)$$

This equation exhibits an inherent ambiguity of 180° in determination of phase by this method. A change of 180° in ϕ does not alter I_{av}, whatever the initial value of ϕ. The orientation of the electric field is also subject to an ambiguity of 180°. Moreover, a change in orientation of 90° can be compensated by a change of 90° in phase, so a new uncertainty is introduced here. In most practical cases there will be no difficulty in resolving this ambiguity in phase on physical grounds.

Experimental verification of eqn. (13) is given in Section 5. However, we shall first describe some of the design details which have been found important in practice.

(4) PRACTICAL DESIGN DETAILS

(4.1) The Waveguide System

Although in the preceding analysis we have assumed for simplicity that the waveguide components are perfectly matched, it is still possible to make measurements of relative phase and amplitude of an electromagnetic field if only the phase-shifter and horn (or other radiator under test) are matched to the waveguide. If these two components are reflectionless, at all positions of the phase-shifter, the reference field at the crystal will be constant for all settings of the phase-shifter, and the field entering the waveguide at the horn will always undergo the same changes in amplitude and phase *en route* to the crystal (apart from phase changes introduced by the phase-shifter itself). The equipment will therefore still measure changes in phase and amplitude of the field at the dipole, and all phases and amplitudes can be referred to the phase and amplitude at some chosen point in the field. The phase-shifter was constructed according to the design given by Halford,[6] scaled down for standard 3 cm waveguide, 0·9 in × 0·4 in internal dimensions, and its standing-wave ratio was better than about 0·98 over the working range. In the actual equipment, the hybrid T was well matched, but, in fact, all measurements were made in the relative way described above.

(4.2) The Dipole and its Supports

In the prototype field-measuring equipment the dipole is mounted at the centre of a length of thin nylon cord, each end of which is rotated by a magslip follower; the magslips are supplied from a Selsyn transmitter driven through a reduction gear from a small motor. It is necessary to drive the nylon at both ends to ensure steady rotation of the dipole.

The magslips are mounted on a U-shaped bracket which is capable of controlled movement in three perpendicular directions (see Fig. 6), and the metallic object nearest to the dipole is at least 10λ distant. No trouble due to reflections from the supports has been noticed.

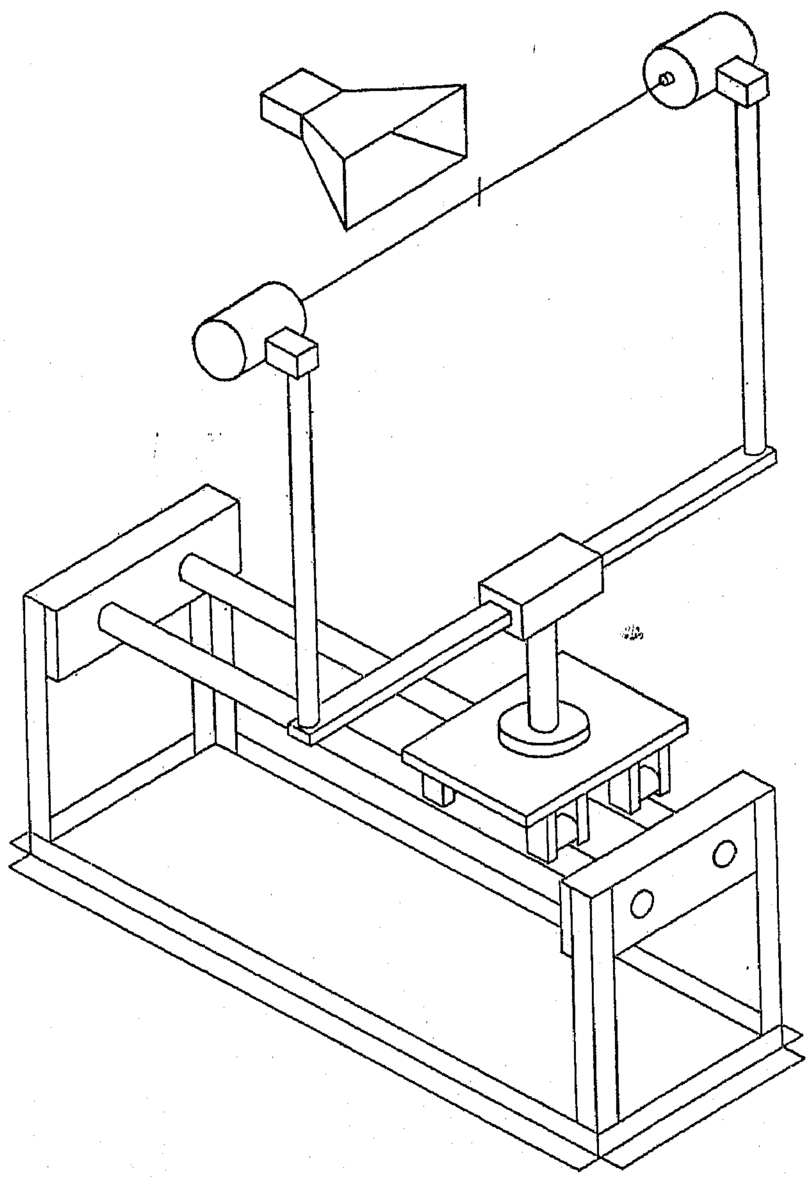


Fig. 6.—Arrangements for rotating and controlling the position of the scattering rod.

The dipole itself is conveniently tuned to resonance (just less than ½λ long) to produce maximum reflected signal, but in regions of strong field or in regions close to metallic objects (e.g. near the mouth of a waveguide horn) the dipole must be reduced in size so as to minimize the effect of multiple reflections between the dipole and the adjacent metal. Since great sensitivity is not usually required in such regions, there is no objection to using a dipole which is a very small fraction of a wavelength long. For example, in order to measure the field across the open end of a waveguide horn (about 1·2 cm × 1·5 cm) the dipole used was 0·02 cm in diameter and 0·18 cm long (roughly λ/18); a slightly larger dipole caused inter-reflection which was just detectable. For a free-space wavelength of 3·2 cm the resonant dipole is approximately 1·4 cm long; a dipole 1·17 cm long will produce about half the response of the tuned dipole.

(4.3) Choice of Commutator Speed

The commutator for phase-sensitive detection is mounted on the shaft of the Selsyn which drives the magslips; the angular position of the commutator brush-gear (which must be known in order to determine the direction of the field) is shown on a calibrated scale.

A disadvantage of a simple rotating commutator is that it rectifies not only the frequency corresponding to its fundamental speed of rotation, but also all the odd harmonics of this frequency. The relative amplitudes of these harmonics are reduced by a factor proportional to their frequencies, but even so, if a periodic interference signal (such as stray pick-up of 50 c/s) of frequency very close to an odd harmonic of the commutator frequency is present, an oscillation of long period will appear on the final indicating instrument. One method of avoiding this difficulty would be to use a high speed of rotation of the dipole, and therefore of the commutator. In practice, sufficiently high speeds cannot readily be achieved, partly owing to the difficulty of making a suitable commutator and partly because the rotating dipole becomes unstable above a certain speed of rotation, in such a way that its centre describes a circle instead of remaining fixed in space. When about 100 cm of nylon was used under about one-half of its breaking strain, a rough test showed that a $\lambda/2$ dipole was unstable in this way if the speed exceeded about 1 500 r.p.m. (corresponding to a commutation frequency of 50 c/s, which is in any case unsuitable). At first sight a convenient speed would appear to be of the order of 750 r.p.m., since the corresponding frequencies rectified by the commutator (25, 75, 125, . . . c/s) all differ by at least 25 c/s from any harmonics of 50 c/s. In spite of the apparent convenience of this particular speed (i.e. the possibility of using a synchronous motor with suitable gear reduction), small errors in alignment of the commutator make it undesirable to run at any submultiple of the mains frequency or its harmonics. If, for example, the brushes (or the commutator segments) are 2° out of alignment, the commutation cycle will be divided into two parts of 179° and 181° instead of two equal parts of 180°, the angular error will be equivalent to 12° at the sixth harmonic and a large proportion of any interference signal at this frequency (150 c/s if the rotational speed is 750 r.p.m.) will be rectified.

Calculation shows that, taking all these restrictions into account, there are three suitable speeds of rotation: 1 050, 825, and 690 r.p.m. (corresponding to 35, 27½ and 23 c/s), each with a tolerance of about 50 r.p.m. (1·7 c/s). The rotational speed of the dipole in the prototype equipment is 840 r.p.m. (28 c/s).

The design of the commutator is an important factor in the satisfactory operation of the apparatus. Two complete spinning-dipole equipments are in use at University College. In the original apparatus a disc-type commutator was used; experience gained with this apparatus suggested that a drum-type commutator would be more satisfactory, and the second apparatus uses a commutator of this kind. The experimental results in Figs. 7, 8, 9, and 11 were obtained with this improved apparatus, since the readings given by it are somewhat steadier.

(4.4) The Amplifier

The maximum gain to be provided by the amplifier depends on the weakest signal to be measured, which, in turn, is limited by the noise and interference level at the input and on the sensitivity of the meter used. A typical value for the standing current in the load resistor (10 kilohms) is 1 mA, and the peak value of the 30 c/s signal varies from about 0·005 to about 5 μA. Thus for the weaker signal the peak value of the 30 c/s voltage is 50 μV. In most of the experiments described the meter used had a full-scale deflection of 100 μA. The resistance of the associated d.c. circuit is about 2 kilohms, so that, allowing for the $2/\pi$ factor introduced by rectification, the voltage needed to produce full-scale deflection is 0·13 volt. Thus a gain of about 2 600 is needed. Provision is made for adjustment of this gain by a stepped voltage divider. The gain of the amplifier and the phase-shift introduced by it must not vary appreciably within the small range of frequencies associated with slight variations in the speed of the driving motor.

The output from the commutator is unidirectional but pulsating, and the meter used to measure the mean current is chosen to have a sufficiently long response time to smooth out these fluctuations.

(5) EXPERIMENTAL RESULTS

Inspection of eqn. (13) shows that the average rectified current I_{av}, varies sinusoidally with ϕ, the phase-shifter setting, and with ψ, the commutator brush setting, other things remaining constant in each case.

Fig. 7 is a plot of galvanometer deflection against phase-shifter setting, and is seen to have the required sinusoidal form. Fig. 8

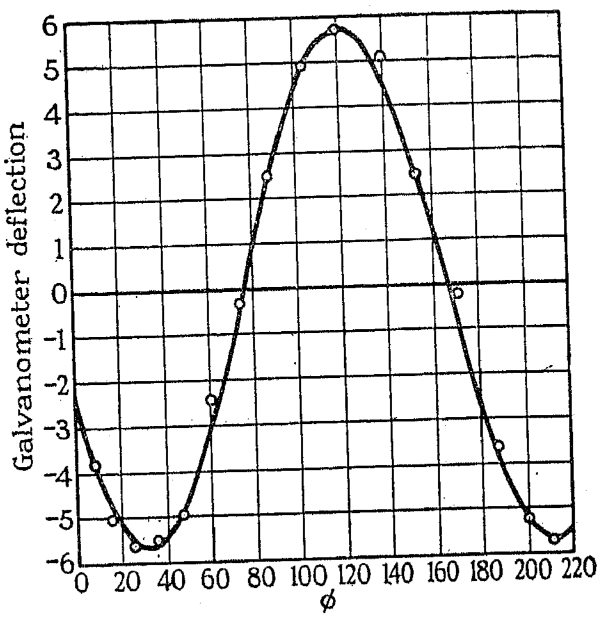


Fig. 7.—Variation of galvanometer current with phase-shifter setting, and cosine curve for comparison.
——— $-5\cdot7\cos 2(16° - \phi)$.
○○○ Experimental points.

is a plot of galvanometer deflection against commutator brush setting, and is also sinusoidal in form. The abscissa is taken as $\psi - 1°$ in Fig. 8, owing to a 1° error in fixing the scale from which the commutator brush position was determined.

The next step is to confirm that the apparatus gives sensible results for field distributions for some simple special cases; we may then have confidence in using it in other cases. To this end, experiments were made using an open-ended waveguide radiator. In the first of these I_{av} was recorded as a function of the distance x of the dipole from the open end of the waveguide. Since $\phi = -kx$, where $k = 2\pi/\lambda$, inspection of eqn. (13) shows that I_{av} should be a periodic function of x, passing through a complete cycle when x increases by half the free-space wavelength.

Experimental results plotted in Fig. 9 confirm this deduction. Next, to illustrate the use of the apparatus for field measurement in a case where simultaneous changes in phase, amplitude and direction of polarization occur from point to point in the field being measured, the arrangement shown in Fig. 10 was constructed. The dipole was arranged to move in a direction perpendicular to the axis of the waveguide in a plane containing the electric vector, i.e. along the line AB in Fig. 10. For each position of the dipole the phase-shifter and the commutator

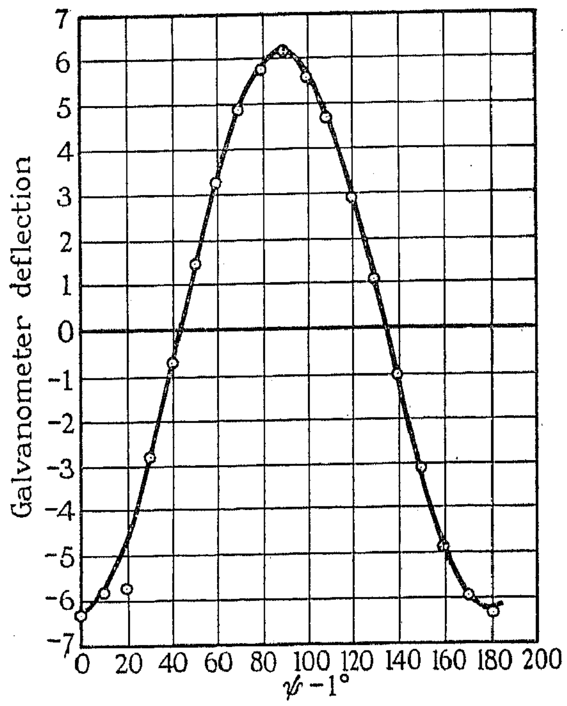


Fig. 8.—Variation of galvanometer current with brush setting, and cosine curve for comparison.
——— $-6 \cdot 2 \cos 2\psi$
○○○ Experimental points.

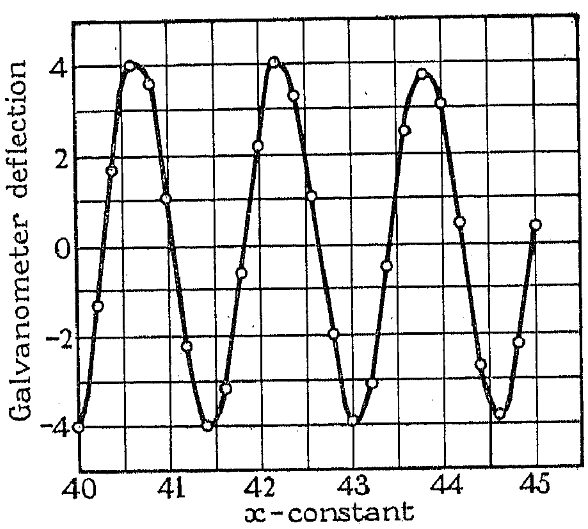


Fig. 9.—Variation of galvanometer current with dipole position, phase-shifter and brush setting remaining fixed.

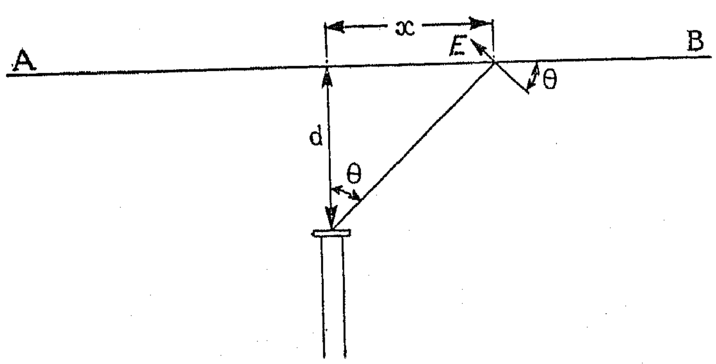


Fig. 10.—Experimental arrangement for simultaneous amplitude, phase and orientation change with change of position.

brushes were adjusted to obtain a maximum value of I_{av}, and the values of ϕ, ψ and I_{av} were recorded.

If measurements are made at a sufficient distance from the radiator for the induction field to be negligible and for the concept of polar diagram to be applicable, the theoretical values of ϕ_t, θ and $|F_t|$ relative to their values at $x = 0$, should be given in terms of x by:

$$\left.\begin{aligned} \phi_t &= k\sqrt{(d^2 + x^2)} - kd \\ \theta &= \arc\tan \frac{x}{d} \\ |F_t|^2 &= \frac{d^2}{d^2 + x^2} f^2(\theta) \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad (14)$$

where $f(\theta)$ is the polar diagram of the radiator. In Fig. 11 the equations for ϕ_t and θ are plotted as functions of x. Points obtained experimentally are plotted on the same graphs, and

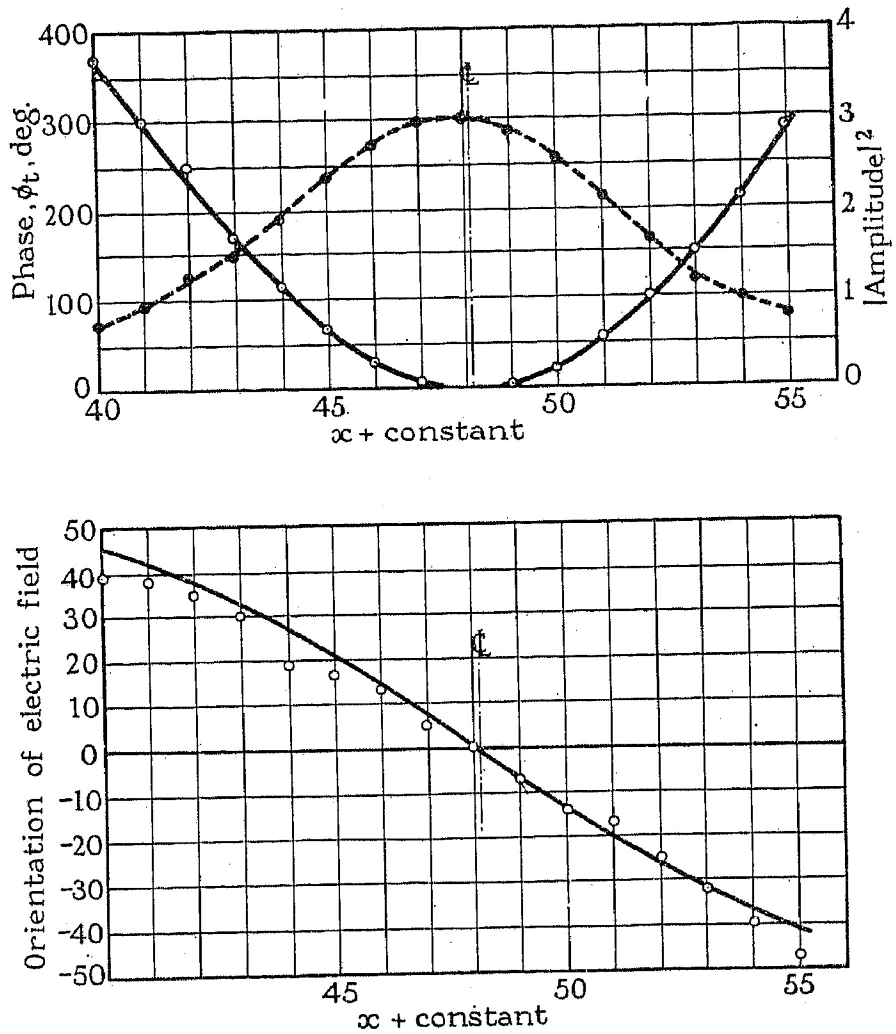


Fig. 11.—Variation of phase, amplitude and orientation of electric field with dipole position.
——— Theoretical curves.
○○ ●● } Experimental points.

agree very well with the theoretical curves. Apart from a single point at $x = 42$ cm, where there is a phase discrepancy of 15° (perhaps due to an error of observation), it appears that measured and calculated phases agree to within about 2°. Orientation of the electric field agrees with the theoretical curve to within about 7° in the worst case; usually the agreement is within 2–3°. Bracketing would reduce random errors in both these sets of figures; the experimental points given here were obtained by a single maximizing adjustment only. The experimental amplitude curve which is also plotted in Fig. 11 cannot be compared with theory, since the polar-diagram function, $f(\theta)$, for the flanged waveguide radiator is not known. It will be noticed that it is not quite symmetrical; this is because the axis of the radiating guide was not quite vertical. (This does not appreciably affect the ϕ_t and θ values.) It is not easy to assess the accuracy of amplitude measurement, but it is probably not better than $\pm 1\%$, and not worse than $\pm 5\%$, of full-scale deflection.

As a final example of the kind of measurement which can be made with this apparatus, Fig. 12 shows contours of equal amplitude near a radiating flanged waveguide having an inductive

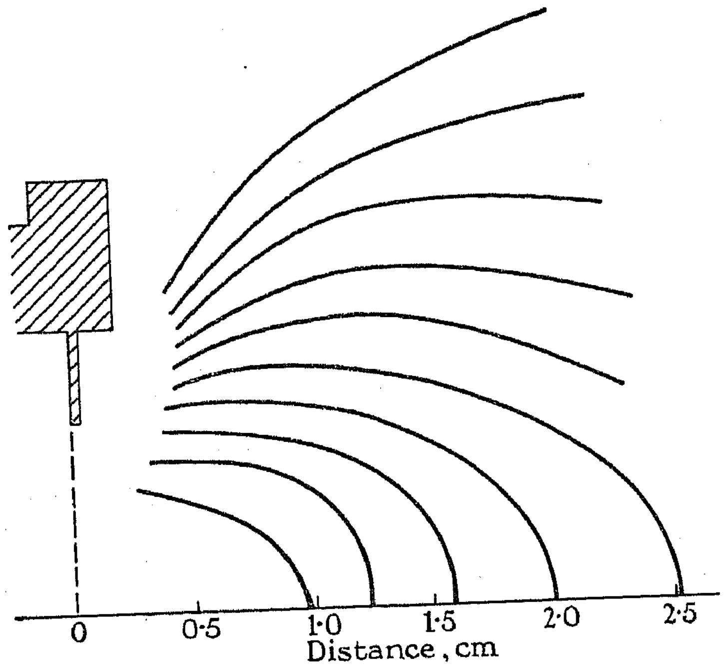


Fig. 12.—Amplitude contours near flanged waveguide radiator: contours approximately 1·6 dB apart.

matching diaphragm near its open end. These contours were found by taking a number of transverse runs across the guide at different distances from the aperture.

(6) THE GENERAL CASE: ELLIPTICAL POLARIZATION

In Section 2 we introduced a simplifying assumption, namely that the field to be measured is linearly polarized in the transverse plane. (Any component perpendicular to this plane, and therefore always perpendicular to the spinning dipole, is, of course, not measured.) We now go back and generalize our formulae to meet cases in which this assumption cannot be made. In the first place, we are now dealing with a field in which E_x and E_y have different phase angles. Thus, the function $F_t(x, y, z)$ must be written

$$F_t(x, y, z) = \boldsymbol{i}|F_x|\varepsilon^{j\phi_x} + \boldsymbol{j}|F_y|\varepsilon^{j\phi_y}) \quad . \quad . \quad . \quad . \quad (15)$$

where F_x, F_y, ϕ_x and ϕ_y are each functions of x, y and z. Referring to Fig. 2, the unit vector $\boldsymbol{u}$ can be written

$$\boldsymbol{u} = \boldsymbol{i}\cos\Omega t + \boldsymbol{j}\sin\Omega t \quad . \quad . \quad . \quad . \quad . \quad . \quad (16)$$

Combining eqns. (15) and (16), we find for the product $\boldsymbol{u}.\boldsymbol{F}$ the result

$$\boldsymbol{u}\,.\,\boldsymbol{F} = |F_x|\varepsilon^{j\phi_x}\cos\Omega t + |F_y|\varepsilon^{j\phi_y}\sin\Omega t$$

Substituting this value in eqn. (2) gives for the backward wave

$$B = \left(\frac{j\omega\alpha Z_0}{ab}\right)A[|F_x|^2\varepsilon^{j2\phi_x}\cos^2\Omega t + |F_y|^2\varepsilon^{j2\phi_y}\sin\Omega t + 2|F_x||F_y|\varepsilon^{j(\phi_x+\phi_y)}\sin\Omega t\cos\Omega t] \quad . \quad (17)$$

We can now use eqn. (9) with eqn. (17) to find the double-frequency component of the rectified current, and the average current flowing in the galvanometer after commutation is then found to be

$$I_{av} = G_1 I_0\{[|F_x|^2\cos 2(\phi_x - \phi) - |F_y|^2\cos 2(\phi_y - \phi)]\cos 2\psi + 2|F_x||F_y|\cos(\phi_x + \phi_y - 2\phi)\sin 2\psi\} \quad . \quad . \quad . \quad . \quad (18)$$

where
$$G_1 = \frac{2GR_1}{\pi R_2}\left(\frac{\omega Z_0|\alpha|}{ab|\rho_0|}\right) \quad . \quad . \quad . \quad . \quad . \quad (19)$$

(It is easily verified that this formula reduces to eqn. (13) for the special case of linear polarization if we put $|F_x| = |F_t|\cos\theta$, $|F_y| = |F_t|\sin\theta$ and $\phi_x = \phi_y = \phi_t$).

To evaluate the four quantities $|F_x|$, $|F_y|$, ϕ_x and ϕ_y, which define a given electromagnetic field at a specified point of observation, it will clearly be necessary to make four observations with different phase-shifts and commutator brush settings.

However, there is an inherent ambiguity in the method (similar to that referred to at the end of Section 3), which can be appreciated by inspection of eqn. (18). If $|F_x|$ and $|F_y|$ are interchanged and ϕ_x and ϕ_y are increased and decreased by $\pi/2$ respectively, there is no change in I_{av} for any arbitrarily chosen phase-shifter and commutation settings. Thus the two different electromagnetic fields corresponding to this interchange cannot be distinguished by the apparatus. In practice, however, the ambiguity can sometimes be eliminated by an *a priori* knowledge of the approximate field distribution. In such circumstances the ambiguity mentioned above is not serious.

The following experimental procedure could be followed.

(*a*) Set brushes to make $2\psi = \pi/2$ and note phase-shifter setting, ϕ_A, for which the galvanometer current reaches a positive maximum value. From eqn. (18), we find

$$\phi_A = \tfrac{1}{2}(\phi_x + \phi_y) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (20)$$

(*b*) Note the value of this maximum current I_A. Using eqn. (18) we find

$$I_A = G_1 I_0 2|F_x||F_y| \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (21)$$

(*c*) Set brushes to make $2\psi = 0$ and leave phase-shifter set to ϕ_A. Then the rectified current I_B under these conditions is given by

$$I_B = G_1 I_0[|F_x|^2 - |F_y|^2]\cos(\phi_x - \phi_y) \quad . \quad . \quad . \quad (22)$$

(*d*) With brushes set to $2\psi = 0$ as above set phase shifter to $\phi_A + \pi/4$. Then

$$I_c = G_1 I_0[|F_x|^2 + |F_y|^2]\sin(\phi_x - \phi_y) \quad . \quad . \quad . \quad (23)$$

From these four relationships it is possible to deduce $|F_x|$, $|F_y|$, ϕ_x and ϕ_y, subject to the ambiguities being resolved by some additional knowledge of the field.

Some preliminary work has been done on this subject experimentally, but it is thought that the method is more likely to be useful for linearly polarized fields for which the ambiguity is more easily resolved.

(7) ACKNOWLEDGMENTS

The authors wish to thank Prof. H. E. M. Barlow for his interest in this work and for providing laboratory facilities. They are also particularly grateful to Mr. P. A. Matthews who constructed the improved model of the original apparatus; the most important improvement was in the design of the commutator—a drum-type commutator of about 1 in diameter. They are also indebted to Mr. J. Shayler for suggesting the use of phase-sensitive detection.

One of the authors (J. C. Parr) is indebted to D.S.I.R. for a grant which made it possible for him to undertake this work.

(8) REFERENCES

(1) Owen Harries, J. H.: "Cavity Resonators and Electron Beams," *Wireless Engineer*, May, 1947, p. 135.

(2) Hansen, W. W., and Post, R. W.: "On the Measurement of Cavity Impedance," *Journal of Applied Physics*, 1948, **19**, p. 1059.

(3) Casimir, H. B. G.: "On the Theory of Electromagnetic Waves in Resonant Cavities," *Philips Research Reports*, 1951, **6**, p. 162.

(4) Maier, L. C., and Slater, J. C.: "Field-Strength Measurements in Resonant Cavities," *Journal of Applied Physics*, 1952, **23**, p. 68.

(5) Midgley, D.: "An Investigation into Microwave Reradiation Problems," Ph.D. Thesis in Electrical Engineering in the University of Leeds.

(6) Halford, G. J.: "A Wide-Band Waveguide Phase-Shifter," *Proceedings I.E.E.*, Paper No. 1466 R, May, 1953 (**100**, Part III, p. 117).

(7) Goubau, G.: "On the Excitation of Surface Waves," *Proceedings of the Institute of Radio Engineers*, 1952, **40**, p. 865.

(9) APPENDIX: FUNDAMENTAL THEORY OF THE METHOD OF MEASUREMENT

(9.1) Application of the Lorentz Reciprocity Formula

In Fig. 13 S_I represents the surface of an infinitely large sphere enclosing the whole apparatus. S_H is a surface containing the transmitter and transmitting horn; this surface can be split into

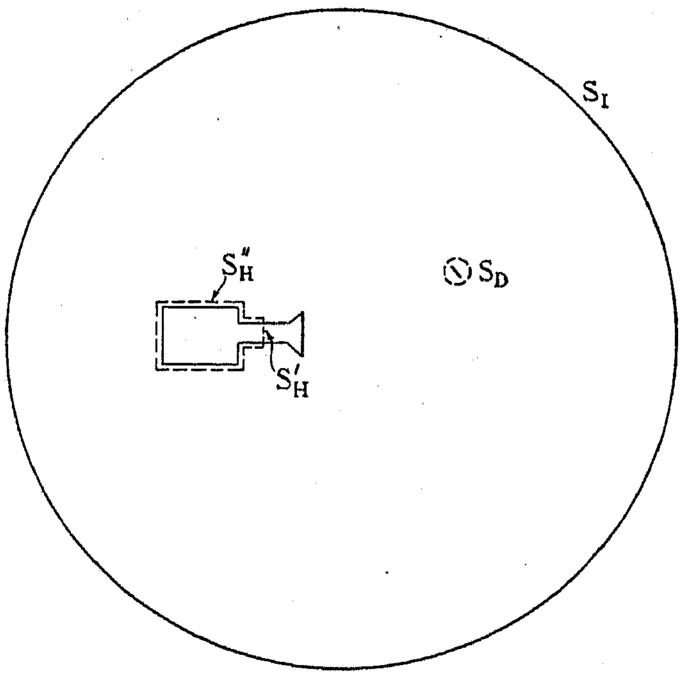


Fig. 13.—Illustrating the analysis.

two separate surfaces, S'_H and S''_H, as shown. S_D is a small sphere surrounding the dipole, which we shall assume to be infinitesimal.

We wish to apply the Lorentz reciprocity theorem in the form

$$\int_S (\ E \times H_2).\ dS = \int_S (E_2 \times H_1).dS \quad . \quad . \quad (24)$$

Eqn. (24) must hold between any two fields E_1, H_1 and E_2, H_2 which satisfy Maxwell's equations everywhere in the volume V bounded by the three surfaces, i.e. in the volume inside S_I and outside S_H and S_D. We choose E_1, H_1 to represent the field produced within V by a travelling wave going outwards through S'_H; the horn is assumed to be matched to the waveguide, so that in the absence of external sources or scattering agencies there will be no wave coming in through S'_H. The field E_2, H_2 is taken to represent the field everywhere within V due to an infinitesimal dipole, of vector dipole moment M, situated at the centre of the small spherical surface S_D. This field includes the field at S'_H produced by the dipole; if the transmitter is matched to the waveguide this will be an inward-travelling wave and there will be no reflected outgoing wave.

We shall obtain a connection between M and E_1 through the electric polarizability of the dipole, and ultimately we shall find the amplitude of the wave reflected back into the waveguide by a dipole of known polarizability.

(9.2) Evaluation of the Integrals

We first divide the surface integrals in eqn. (24) to give

$$\int_{S_I} (E_1 \times H_2).dS + \int_{S_H} + \int_{S_D} = \int_{S_I} (E_2 \times H_1).dS + \int_{S_H} + \int_{S_D} \quad . \quad . \quad . \quad . \quad (25)$$

In the following analysis these integrals will be considered in turn.

(9.2.1) The Surface S_I.

On the surface S_I we have

$$E_1 = -\,n \times H_1\sqrt{\frac{\mu_0}{\epsilon_0}}$$

$$E_2 = -\,n \times H_2\sqrt{\frac{\mu_0}{\epsilon_0}}$$

where n is an outward unit normal. By writing the surface integrals in terms of H alone from these relationships, the scalar triple products $(n \times H_1) \times H_2 . n$ and $(n \times H_2) \times H_1 . n$ occur; these can both be written in the symmetrical form $-(n \times H_1) . (n \times H_2)$, from which it is clear that:

$$\int_{S_I} (E_1 \times H_2).dS = \int_{S_I} (E_2 \times H_1).dS \quad . \quad . \quad (26)$$

(9.2.2) The Surface S_H.

The integrals over S''_H vanish. For example, E_1 is parallel with the normal to S''_H everywhere, since S''_H is assumed to be a perfectly conducting surface, and E_1 can therefore be written $E_1 = nE_1$. Then

$$(E_1 \times H_2).dS = (E_1 \times H_2).ndS$$
$$= E_1(n \times H_2).ndS = E_1(n \times n).H_2dS = 0$$

and similarly for $(E_2 \times H_1).dS$.

Evaluation of the integrals over S'_H involves consideration of the field distribution inside the waveguide. We shall assume that only the dominant mode can be excited, and—to avoid duplication—that this is an H-mode.

Let the axial co-ordinate in the waveguide be x, so that the outward normal on S'_H is $-i$. Then the fields associated with an outward-travelling wave at S'_H are

$$E_{tan};\ H_{tan} = -\,i \times E_{tan}/Z_0;\ E_z = 0 \quad . \quad . \quad (27)$$

where $$Z_0 = \sqrt{\left(\frac{\mu_0}{\epsilon_0}\right)}\frac{\lambda_g}{\lambda}$$

Let us write

$$E_{1tan} = Aw;\ E_{2tan} = Bw \quad . \quad . \quad . \quad . \quad (28)$$

where w is a vector function of position in the transverse plane S'_H. The associated magnetic fields are

$$\left.\begin{aligned} H_{1tan} &= \quad (w \times i)A/Z_0 \\ H_{2tan} &= -\,(w \times i)B/Z_0 \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad (29)$$

The negative sign in the second equation arises because the field E_2, H_2 is an incoming wave, whereas E_1, H_1 is an outgoing wave. Using these equations

$$\int_{S'_H} (E_1 \times H_2).ds = -\,ABY_0\int_{S'_H} w \times (i \times w).idS$$

$$= -\,ABY_0\int_{S'_H} (i \times w).(i \times w)dS$$

$$= -\,ABY_0\int_{S'_H} w^2dS \quad . \quad . \quad . \quad . \quad (30)$$

For the H_{01} mode in rectangular waveguide, $w = j \sin(\pi z/b)$, and we obtain in this case

$$\int_{S'_H} (E_1 \times H_2) . dS = -ABY_0 \frac{ab}{2} \quad . \quad . \quad . \quad (31)$$

and similarly
$$\int_{S'_H} (E_2 \times H_1) . dS = +ABY_0 \frac{ab}{2} \quad . \quad . \quad . \quad (32)$$

(9.2.3) **The Surface S_D.**

Consider first the integral $\int_{S_D} (E_1 \times H_2) . dS$. In this integral dS is directed out of the volume V towards the centre of S_D; hence, applying Gauss's theorem, we have

$$\int_{S_D} (E_1 \times H_2) . dS = -\int_v \text{div}\,(E_1 \times H_2) dv$$

$$= -\int_v (H_2 . \text{curl}\, E_1 - E_1 . \text{curl}\, H_2) dv \quad . \quad (33)$$

where v is the volume within S_D.

Within the volume v, curl $E_1 = -j\omega\mu_0 H_1$ and curl $H_2 = J + j\omega\epsilon_0 E_2$ where J is the current density associated with the oscillating charges of the dipole within. We now obtain

$$\int_{S_D} (E_1 \times H_2) . dS = \int_v (j\omega\mu_0 H_1 . H_2 + j\omega\epsilon_0 E_1 . E_2 - E_1 . J) dv$$
$$. \quad . \quad . \quad . \quad . \quad (34)$$

Since curl $H_1 = j\omega\epsilon_0 E_1$, with no associated current flow, we also have the similar equation

$$\int_S (E_2 \times H_1) . dS = \int_v (j\omega\mu_0 H_1 . H_2 + j\omega\epsilon_0 E_1 . E_2) dv \quad . \quad (35)$$

and combining eqns. (34) and (35), we have

$$\int_{S_D} (E_1 \times H_2 - E_2 \times H_1) . dS = \int_v E_1 . J dv$$

If the current flow J is associated with a thin wire whose axis is parallel to the unit vector u, and if E_1 is virtually constant within the small volume v, we have

$$\int_v (E_1 . J) dv = u . E_1 \int J dv = u . E_1(Il) = j\omega u . E_1(ql) \quad . \quad (36)$$

But $u(ql)$ is the electric dipole moment, M, and so

$$\int_{S_D} (E_1 \times H_2 - E_2 \times H_1) . dS = +j\omega M . E_1 \quad . \quad (37)$$

This equation has been obtained previously by Goubau,[5] but as the derivation is not given in Goubau's paper it is thought worth while to provide a proof here, particularly since this is a basic equation upon which the method depends. Thanks are due to Dr. Goubau for a very elegant proof of eqn. (37) using the Hertz potential representation of the dipole field.

(9.3) **The Final Result**

Substituting from eqns. (26), (31), (32), and (37) into eqn. (25), we find that

$$ABY_0 ab = +j\omega M . E_1 \quad . \quad . \quad . \quad . \quad (38)$$

Since E_1 is clearly proportional to A it is convenient to introduce the normalized electric-field distribution function $F(x, y, z)$, giving

$$E_1(x, y, z) = AF(x, y, z) \quad . \quad . \quad . \quad . \quad (39)$$

Here (x, y, z) is the position of the centre of the dipole. Substituting eqn. (39) in eqn (38) gives

$$B = (j\omega Z_0/ab) M . F(xyz) \quad . \quad . \quad . \quad . \quad (40)$$

In the present example, the dipole moment M is produced by the action of the electric field E_1 on a linearly polarizable thin rod according to the equation

$$M = \alpha(u . E_1)u$$

or
$$M = A\alpha(u . F)u \quad . \quad . \quad . \quad . \quad . \quad . \quad (41)$$

where α is the polarizability of the rod. Substituting eqn. (41) in eqn. (40) gives, finally,

$$B = A\left(\frac{j\omega\alpha Z_0}{ab}\right)(u . F)^2 \quad . \quad . \quad . \quad . \quad (42)$$

This is the fundamental formula on which the method depends.

DISCUSSION ON "IMAGE INTENSIFICATION IN RADIOLOGY"

The discussion, which took place before an Extra Meeting of THE INSTITUTION *on the* 12*th May*, 1955, *was opened by W. J. Oosterkamp, M.A., Ph.D., and G. M. Ardran, M.D., summaries of whose introductory remarks are given below.*

CONSIDERATIONS IN IMAGE-INTENSIFIER DESIGN

By W. J. OOSTERKAMP, M.A., Ph.D.

Several systems for image intensification have been put forward in recent years: by means of a scanning X-ray beam; by television pick-up of the image on a fluorescent screen; and by electronic image-intensifying tubes. Only the latter systems, developed independently by Coltman and by Teves and Tol, have become commercially available, and they are being used in increasing numbers in clinical radiology.

Both tubes are fundamentally alike, but there are major differences in design and execution. In the tube developed by Teves and Tol a photocathode is in contact with a fluorescent screen. The incident X-ray image is converted into an image of photo-electrons, which are accelerated to 25 keV and focused on a viewing screen by a single 3-electrode electrostatic lens. The image on the viewing screen is linearly nine times smaller than that on the X-ray screen, but its brightness is at least 1 000 times greater. This image can be viewed through an optical magnifying system or it can be photographed.

The image intensifier has raised the perceptibility to such an extent that fluctuation phenomena, inherent in the quantum nature of X-rays (X-ray noise), have become a limiting factor. An object detail cannot be observed with certainty if the contrast between it and the surrounding area is not at least three times greater than the statistical fluctuations, of the luminance in fluoroscopy or the density in radiography. In communication terms this means that the signal/noise ratio has to be at least three. This ratio, and therefore the theoretical limit of perceptibility, will be shifted to smaller details and smaller contrasts when the X-ray intensity (in fluoroscopy) or the time integral of intensity (in radiography) is increased.

Fluctuations can also be a limiting factor when the image intensifier is not used, e.g. in regular fluoroscopy and miniature radiography. In both instances, however, the limiting fluctuations are much larger than the X-ray quantum fluctuations; owing to the poor optical coupling between the fluorescent screen and the eye of the observer or the photographic film, only one out of every 100 X-ray photons absorbed in the fluorescent screen gives rise to a visual impression in fluoroscopy or to a developed silver grain in miniature radiography. The fluctuation level is reduced to that of the X-rays when the image intensifier is used, for then every X-ray photon absorbed in the screen is registered, by virtue of the intensification.

The geometrical aberrations in the image formation arising from the light diffusion in the fluorescent screen, the width of the focal spot, the movement of the object, etc., also limit the dimensions of the smallest object that can be observed. In order to minimize the X-ray dose to which the patient is exposed, the X-ray intensity should ideally be not higher than is necessary to carry all the information the instrument is able to resolve and display. In other words, the geometrical resolving power of any method of radiography or fluoroscopy should be such that the limit set by the X-ray noise is closely approached. This is more difficult to realize when large amounts of X-rays have to be used, such as in full-size radiography.

The contrast to X-ray noise depends upon the penetrating power (hardness) of the X-rays, which is mainly determined by the voltage across the X-ray tube. The intensity of the X-rays, having passed the salient, increases and the X-ray noise decreases when more penetrating radiation is used at a certain incident intensity. However, more penetrating radiation means smaller contrast. The optimum tube voltage has been calculated, assuming a wavelength distribution of the radiation according to the Kulenkampff–Kramers formulae, and the decrease of the X-ray absorption in the fluorescent screen for harder radiation has been taken into account. The reduction of contrast by the scattered radiation is a complicating factor, and this effect becomes more difficult to eliminate as the radiation becomes more penetrating.

SOME PROBLEMS IN MEDICAL APPLICATIONS

By G. M. ARDRAN, M.D.

In medical fluoroscopy, because of the radiation hazard to both patient and operator, the fluorescent screen is used at very low visual brightness. By causing the X-ray image to produce an electron image which can then be intensified, the visual brightness may be effectively increased 200–1 000 times. It is not possible to reduce the X-ray intensity below about 10% of that usually employed, because statistical fluctuation of the X-ray quanta at these low intensities reduces the resolution. Thus the intensifier now gives the choice between a brighter image of better definition or a reduction in the radiation dose. It is frequently not possible to take full advantage of the intensifier in clinical practice, owing to the small size of the field covered and the relatively cumbersome nature of the apparatus.

The intensifier has also enabled us to make cinematograph films with about 10% of the radiation dose required for a full-sized film taken with intensifying screens, or about 1% of that previously required for photofluorography. The definition is usually better than that obtained in conventional photofluorography, but never as good as that obtained with a full-sized film. As an indication (easily appreciated in a clinical department) of the definition obtained with the intensifier it can be stated that in our experience it is not possible to see the lines caused by a stationary 6 : 1 Lysholm grid; these can always be seen on the full-sized film. Thus the image intensifier is of value for making radiographs only when low dose is of importance and when definition is not critical, or when an image reduced in size is required for cinematograph work. Apart from its value in fluoroscopy, the main virtue of the intensifier lies in the compara-

tive ease with which cinematograph film can be made. Such films are at present valuable for research, and they are of undoubted value in the clinical investigation of swallowing processes in the mouth, pharynx and upper oesophagus; but it remains to be seen how valuable they are for routine clinical diagnosis in other fields. When cinematograph films are being made it must be remembered that, although the dose per picture is relatively small, the total dose may soon exceed desirable limits if care is not taken. Now that the abdomen and pelvis can be examined by cineradiography, this may easily lead to an excessive dose to the gonads.

DISCUSSION

Dr. A. Nemet: I should like Dr. Ardran's comments on the alternative methods of angiography and cineradiography from the aspect of skull examination, e.g. for brain tumours. The advantages of cineradiography are low dosage and more pictures per second to record the fast movement of the dye at the various phases of its travel; on the other hand, angiography gives better definition.

In medicine a compromise must be found between low dosage and good definition; in industrial fluoroscopy and radiography it must be between economy and definition. The high intensification factor and good screen definition of the image intensifier allows the compromise to be reached at a higher quality of screen image than is possible with ordinary fluoroscopy. We have carried out laboratory tests on both fluoroscopy and radiography, using wire penetrameters and expressing the sensitivity as the ratio of the wire diameter against the plate thickness. Figs. 1(*a*) and 1(*b*) compare the results obtained with and

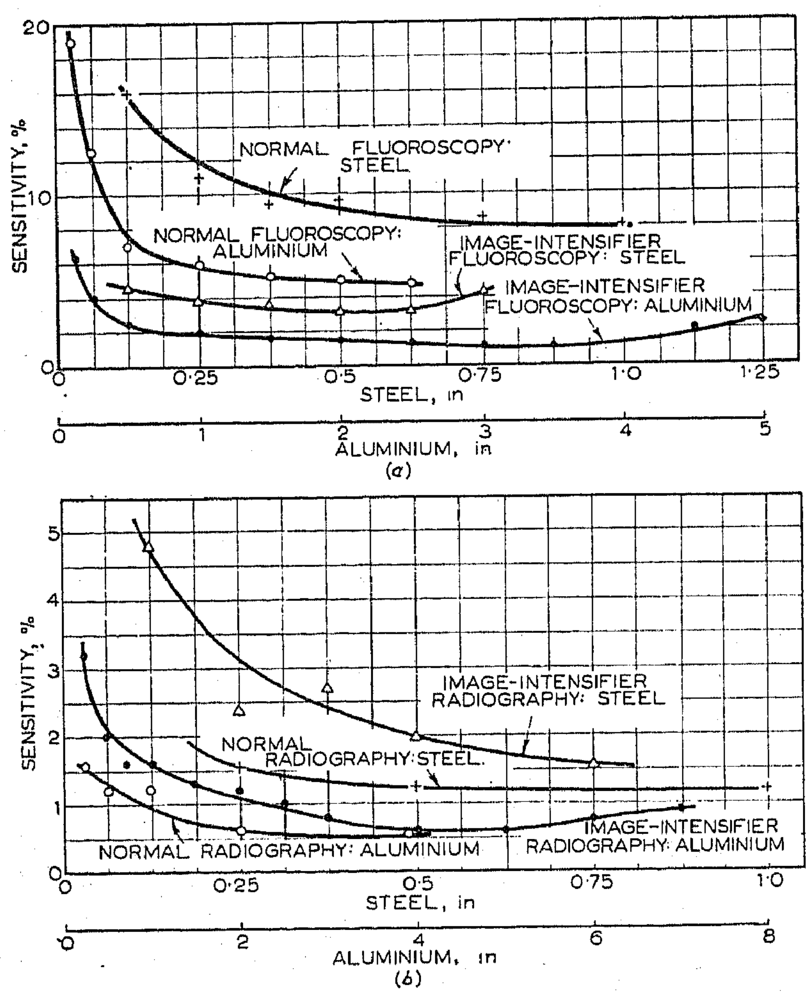


Fig. 1.—Variation of sensitivity (D.I.N. penetrameter) with thickness for normal and image-intensifier fluoroscopy (*a*) and radiography (*b*) of steel and aluminium.

without image intensification for various thicknesses of steel and aluminium. It will be seen that the image intensifier improves the sensitivity from about 8–10% to about 3–5% for fluoroscopy of steel in the thickness range ¼–1 in; the improvement for aluminium is even more marked, normal fluoroscopy giving 5–8% over the range ½–2½ in and the image intensifier 1–2% for thicknesses up to 5 in—an attractive approach to radiographic values. For radiography [Fig. 1(*b*)] the position is reversed, although the difference is less marked; image intensification reduces the sensitivity from about 1¼% to 1½–3% for ¼–¾ in steel and from 0·5–1% to 0·5–1·5% for 1–4 in aluminium.

Thus image-intensifier fluoroscopy gives results between normal fluoroscopy and normal radiography, tending towards the latter, while intensifier radiography gives sensitivities about the same or a little better. These results may reveal uses of the fluoroscopic method which are impossible with conventional fluoroscopy, because of its poor quality, and where normal radiography would be too expensive. On the other hand, intensifier radiography is important where a permanent record is necessary. A further attractive feature is the large tube voltage reduction permitted by the image intensifier.

Some comparison of the visual results are shown in Figs. 2 and 3. Fig. 2 compares normal and image-intensifier radiographs of an aluminium casting; the flaws are visible in both, but the different penetration in the various sections is quite marked. Fig. 3 shows similar results for a weld between 0·15 in

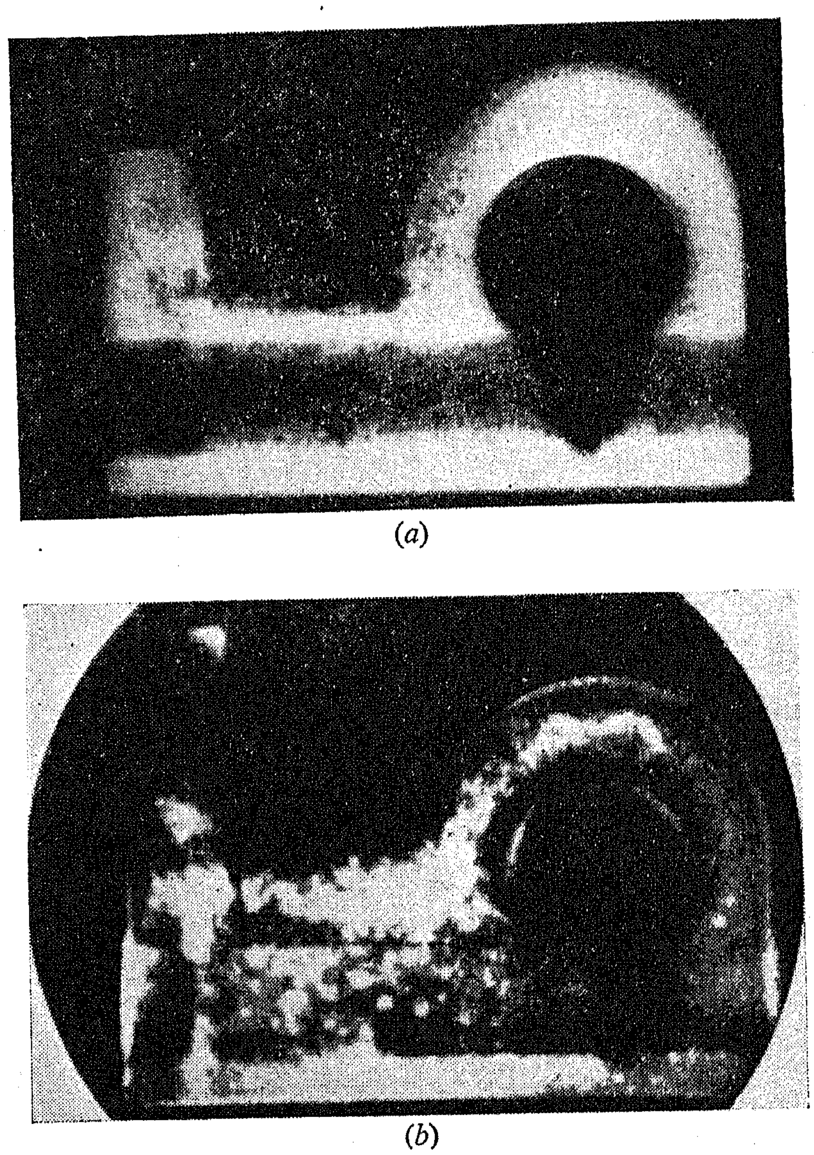
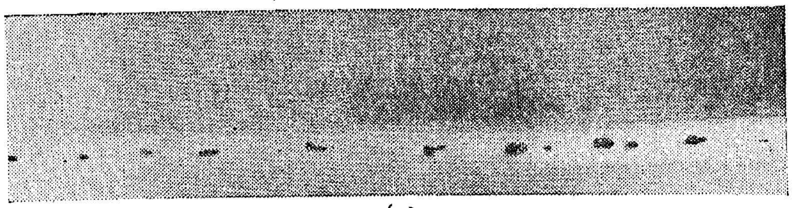
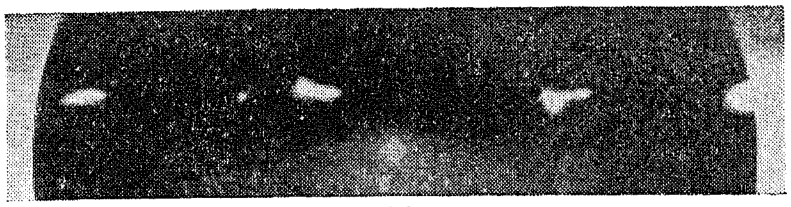


Fig. 2.—Normal and image-intensifier radiographs of an aluminium casting.
(*a*) Normal.
(*b*) Image intensifier.



(a)



(b)

Fig. 3.—Normal and image-intensifier radiographs of a steel weld.

(a) Normal.
(b) Image intensifier.

steel plates; the slight crack is visible in both. In these illustrations the normal radiograph is negative and the image-intensifier one positive; exposure data for normal radiography are 80–100 kV, 10 mA and 4–5 min, while for image-intensifier work they are 40–60 kV, 4 mA and 40 sec.

Dr. R. Astley: At the Birmingham Children's Hospital we have two main uses for the image intensifier: in fluoroscopy when room light is needed to handle surgical instruments, e.g. in cardiac catheterization; and in angiocardiography, where the rapid events in the passage of an opaque fluid through the heart must be recorded. Here the intensified image is photographed at 32 frames/sec on a fine-grain slow-speed 16 mm film (perhaps thereby avoiding some of the quantum effects which have been mentioned). The ciné camera is linked to the X-ray generator, so that each frame is exposed as an individual radiograph; this halves the dosage to the patient and permits higher tube currents and lower tube voltages, thereby aiding the attainment of sufficient contrast—one of the main problems.

The improvements I should like in the image intensifier are better definition, increased image contrast and a larger field, the first being the most important. Lack of sharpness is often not objectionable when the film is projected at the normal rate, but it becomes very apparent when the film speed is reduced or projected frame by frame, as is usually necessary for analytical purposes.

Mr. S. B. Osborn: Until they go into the screening room and see for themselves, few people appreciate just how little light normal fluoroscopy makes available to the radiologist for diagnostic purposes, and anything which could be done to increase the amount would be very welcome.

I am glad to note the use of the term "image intensification"; the often-used "image amplification" is a misnomer, because the image produced is smaller than the original.

One can make use of the increase in brightness for a given amount of radiation either to reduce the dosage to the patient or to increase the visuality for, and so the ease of, diagnosis. The former is obviously preferable, because large quantities of radiation are used in fluoroscopic diagnosis; the clinical radiation worker is restricted to a maximum dosage of 0·3 röntgen per week, but it is not unusual for a patient to receive 100–200 times this amount during a single series of examinations on one day. There would indeed seem little point in worrying about a few milliröntgens of radiation received from the fall-out of an atomic explosion on the other side of the world while we may administer scores of röntgens during diagnostic examination.

I would add to the pleas for a screen larger than the present 5 in-diameter one, which imposes considerable restrictions; many of us are accustomed to fluorescent screens about 14 in across.

Image intensification might help medical radiology from another aspect, in that the X-ray tube might be replaced by radioactive sources when we can make better use of the small amounts of radiation. Soft γ-ray emitters have been used in the United Kingdom, and β-particle emitters have been used in America as a source of X-radiation. There may well be a future for this type of radiology as the right sources become increasingly available.

Mr. R. Halmshaw: In industrial radiology we have no problems arising from the effect of the dose on the specimen, but we are much concerned with definition. We compare results with radiographs taken with metal, rather than tungstate, screens, and the sharpness may be as small as 0·1 mm. How closely can the image intensifier approach this? There seem to be two limitations to definition—the light diffusion in the primary screen, which, I believe, limits the resolution to about 0·3 mm, and the small size of the second screen, which means that the image must be magnified about ten times for adequate viewing; from the industrial aspect the blurring of these two screens is considerably more important than quantum-fluctuation limitations.

What causes the spurious light spots on the screens of some of these tubes, and what are the possibilities of eliminating them? I gather that the dark spots are dust marks on the screen, but I have seen a diffuse spot of higher intensity at the centre of the screen, which is very confusing when viewing industrial specimens.

I was interested in Dr. Nemet's curves of D.I.N. penetrameter sensitivities, but wonder what he means by "normal" radiography. Were the full-size photographs based on radiography with tungstate screens, or were other screens used? From the industrial aspect there is a whole range of techniques available, and the results differ markedly according to which is used.

Mr. M. F. Osmaston: I am interested in the effects of the various X-ray voltages used for radiology and the application of γ-ray sources to the image intensifier. I imagine that the γ-rays have a greater penetration of the primary screen, so that a thicker screen is necessary and the definition is consequently reduced. Again, with the low rates of quantum arrival which may occur when using γ-ray sources, can the image intensifier be used visually or is the brightness with, say, 6 in thick steel specimens so low that photography is essential?

Is it possible to combine the image intensifier with a television pick-up tube to improve contrast, the photocathode of the latter being in contact with the secondary screen of the former? Besides improving contrast, this might permit presentation of the picture where the doctor found it most convenient. Moreover, a 2-stage tube in which the intermediate screen and photocathode were well matched to each other in light colour output and sensitivity—rather than the phosphor colour being suited to the sensitivity of the human eye—would give a considerable gain.

Mr. R. Meakin: The final image in the intensifier is about a centimetre in diameter, which seems rather a waste of 35 mm film, so that an enlargement of about twice would have much to commend it. However the final image is photographed it will involve an ordinary high-aperture camera lens and a similar supplementary lens. If $f/15$ lenses are used there may be a dozen components, so that, even with coated surfaces the lenses must be responsible for a definite loss of contrast. Have the lens manufacturers been approached about a special lens for this purpose?

I am not clear about the effect of quantum fluctuation on the photography of the image. The quantum limit is rather vague, and estimates of its seriousness vary. It seems to be associated with the low integration time of the eye, so that unless the rate of arrival of quanta is high the standard deviation will be high. This does not apply in photography, since the film must be given

a certain exposure to produce a useful density; thus a quantum fluctuation which might be inadmissible for normal viewing might not produce a sensible effect on film. It has been suggested that the intensification factor might be increased to about 10000 before quantum fluctuation interfered with photography.

Finally, I should like to support Mr. Osborn's plea for a larger screen.

Prof. J. D. McGee: What is the ratio between the effectiveness of the image intensifier when used with maximum sensitivity monocularly and with binocular viewing using mirrors? The latter is considerably less efficient than the former, and I should like some more accurate appraisal of the difference.

Why does Dr. Oosterkamp restrict his voltages to 25–30 kV? Is it not possible to use 50 or 100 kV, when there would obviously be more fluorescent energy?

To what extent is definition limited by the phosphor of the viewing screen? During the war, one of the major problems of the image convertors of this general design was to obtain a viewing-screen phosphor of sufficiently fine particle size not to limit the definition when viewing with a magnification factor of perhaps ten.

Mr. G. R. Airth: Originally I obtained my image intensifier for the examination of babies, and to me its most remarkable property is its definition, for I am not so concerned with anything else. In a new-born baby the outlet of the stomach—the pyloric canal—may be 1·5 mm in diameter. With a normal X-ray screen I can infer its presence because I can see filling of the duodenum distally; I can take photographs in rapid succession, and perhaps one out of a dozen will show an image of three streaks of barium in the canal. With the image intensifier I can see it directly and watch it working, which is a great advance, even if I am not yet able to photograph it.

On the other hand, I find the central white spot most annoying. Initially it was very small, but now it is about 1·5 cm in diameter, and I find it difficult to concentrate on a part of the screen which is well away from the spot but which does not give maximum definition. Admittedly I can accommodate a large part of a baby within the 5 in screen, and so have none of the limitations involved in examining adults; for adult examination I have found a high-ratio grid of advantage; for infants, a 0·3 mm source.

In pelvimetry we normally measure both the pelvis and the baby's head indirectly and then judge how the head will enter the mother's pelvic brim. With the image intensifier there is sufficient light on the screen to perform direct pelvic cephalometry: we can, as it were, fit the baby's head into the brim under direct vision. It is difficult, and there is really very little light; but it can be done and it has never before been possible.

Dr. V. Quittner: I should like Dr. Oosterkamp's views on systems of image intensification other than those now in permanent use. I believe that the Mond system has very little advantage, for it requires a very large X-ray tube and very high currents; moreover, the target of the X-ray tube must absorb a considerable amount of power. On the other hand, the television system developed by Morgan does have advantages, although it is not used at present: it can be reproduced in any size; it can be intensified practically without limit; and it can even be shown in an auditorium. Is there some future for this system?

Dr. W. J. Oosterkamp (*in reply*): There is undoubtedly a big gap between the first realization of a new instrument by the scientist in the laboratory and the final instrument that will be really useful in clinical practice. While the main problems of image intensification have now been solved, there is much work still to be done before it becomes a universally useful tool. In particular, there is the way in which the image should be presented; a doctor works with living material, and must make his investigations quickly, because of the possible harmful effect of the radiation. A final solution to these problems has probably not yet been reached, and further improvements will take time. As the apparatus is now used by the medical profession, further progress will be stimulated by their comments and inquiries.

Experimental tubes with larger fields have been made and films produced with them have been shown, but it will be a long time before they will be available on a production scale, together with the auxiliary equipment necessitated by their use in radiography, cineradiography and viewing. There are no insurmountable obstacles, although bigger tubes will be even more cumbersome than the present ones.

Mr. Osborn asks about the use of radioactive isotopes; this will always be limited by quantum fluctuations. A certain quantity of photons is needed to convey a certain amount of information, and large sources of isotope would have to be used to compete with the small focal spot of a simple X-ray tube.

The blurring of the X-ray screen is the major cause of lack of definition at the moment. It is slightly more than that with normal intensifier tungstate screens and considerably less than that inherent in the normal fluoroscopic screen. The viewing screen itself is made of a very fine powder, and its contribution to blurring is less than that of the X-ray screen. If γ-rays or fairly hard X-rays are used, a thicker primary screen is needed and there will be increased blurring.

The limit to the accelerating voltage is set by the field emission from the cold-cathode surface, but I agree that it could be rather higher than 25 keV. On the other hand, this might make the tubes less stable and the viewing screen would have to be thicker to stop the electrons. The present intensification factor is more than ample, because quantum fluctuations are already playing a part; there is little point in increasing the intensification, for less quanta would be needed to obtain a given result and the fluctuations would consequently become even more apparent. This also applies to the suggestion of a 2-stage tube: there is no need for this, for the intensification level reached already is on the high, rather than the low, side for practical purposes.

Mr. Meakin asks whether quantum fluctuations are also visible in radiography. With a very sensitive emulsion, perhaps 20 or 30 developed grains are the result of one X-ray photon absorbed in the X-ray screen, and then quantum fluctuations are clearly visible, but they can be reduced if less-sensitive film is used or if the light-gathering power of the optical system is reduced. The contrast could indeed be increased considerably by a television system, such as that used by Morgan, but it is impossible to improve the signal/noise ratio of the primary image. No electronic device can improve the signal/noise ratio; therefore the increased contrast may not, under certain circumstances, improve perceptibility when this is limited by the quantum fluctuations. The disadvantage of television systems is that the pick-up tubes and the amplifiers have their own noise background. One advantage of the image intensifier in its present form is that its background current, caused by field emission, is negligible compared with the current at X-ray intensities used in medical radiology. It is a little different in the industrial field, where the X-ray intensities are often much lower. If radiographs are made with exposures of 1 min or more with high-power optical systems, the background of field emission might become a severe handicap.

There is one important point with regard to the improvement of contrast. The intensifier greatly increases the luminance and it approaches the values at which normal radiographs are being viewed. On the photograph the contrast is increased by a factor of 2·5 or even 3, as the characteristic curve of the emulsion is non-linear; but the image intensifier is linear, so that it sometimes appears that contrast is lost in the image intensifier. In fact, very little of the contrast is lost, but the visual impression is

much poorer compared with the radiograph due to its 2·5–3 times greater visual contrast.

I cannot give Dr. McGee the data he requires about binocular vision compared with the mirror system. The binocular microscope is the ideal device from the aspect of presenting all information contained in the image on the viewing screen to the eye, but it is not always convenient. I think that a simple mirror system is a good compromise between convenience and the loss of information for the benefit of the man who is engaged in daily clinical practice. We physicists are sometimes inclined to go for fundamentals and to neglect such practical factors as the presentation of the image to the radiologist who has to work with it.

The white spot mentioned by Mr. Halmskaw and Mr. Airth is due to the ionization of gas in the tube that develops with age. This annoying effect was rather frequent in older tubes but it has been overcome now almost completely.

Dr. G. M. Ardran (*in reply*): I can offer Dr. Nemet no practical experience of cerebral angiography on ciné film, but I have little hope of this procedure being valuable. Some work was done 5–7 years ago in Oxford; full-size films were taken at 2 or 3 frames/sec, and several hundred cases were examined in this manner, but information additional to that from the usual 3- or 4-exposure technique was obtained in only a relatively small number of cases. The procedure is now carried out comparatively rarely.

In cerebral angiography the detail required is very great, whereas in cardioangiography we are dealing with large vessels and chambers and with the presence or absence of contrast medium. The cerebral vessels with which we are concerned may frequently be individual vessels or collections of vessels of the order of a millimetre or less in diameter. Cerebral angiography requires definition rather than a large number of pictures. I feel sure that whenever angiography requires detail, the full-size technique—particularly when a fairly large number of pictures can be taken—will be the method of choice.

For screening purposes, particularly where the tube is relatively close to the patient (usually of the order of 20 in or so), the 0·3 mm focal spot has a use. I use one for screening under these circumstances, and believe the image to be sharper than with a 1·0 mm spot. For direct enlargement a 0·3 mm spot is essential, but if the tube is used at radiographic distances I do not think the focal spot matters very much. One cannot visualize the stationary 6 : 1 Lysholm grid lines with the intensifier—even using a 0·3 mm focal spot—as one can with any size of focal spot and the fastest film-screen combination one can get in the ordinary way.

Enlargement techniques can be very valuable, as industrial radiologists might bear in mind. The lack of definition in the intensifier may to some extent be overcome by an enlargement technique, particularly now that 0·2 and 0·3 mm effective-focal-spot tubes are available with water-cooled anodes for continuous operation.

It has recently been suggested that the quantum limit has not, or should not have, been reached. Not all X-ray quanta can be stopped in fluorescent screens or other recording materials, and not all can be used effectively. This limit seems to have been reached in the intensifying tube using a fluorescent material as the first intercepting level—as it has with the ordinary intensifying-screen combination. The limits within which one can use these two methods for taking a picture are of the same order of magnitude. If a picture taken with the intensifier shows a "grainy" pattern larger than the grain of the film and another picture is taken a fraction of a second later, the grain pattern is of the same order of magnitude but has a different arrangement. The same applies with intensifying screens: if a picture is taken with an intensifying screen with a large dose of radiation, using Microfile film (which requires a large dose of radiation), the film shows a "grainy" pattern which is due to the fluorescent material; another picture taken a few minutes later will show the identical grain pattern. But if one uses a fast film combination and takes pictures at intervals, the grain pattern will be different every time. This point is of great interest to engineers and technical people, since, so far as I can see, it means that the character of ordinary X-ray tubes and units will not change.

We appear to have reached the limit, and I cannot see that any other form of intensification is likely to be a great improvement on the methods we now have. This, I think, answers the question about the use of isotopes: unless we can get an isotope source with the same output as a modern X-ray tube and coming effectively from a source equivalent in size to the focal spot in an X-ray tube, I cannot see that it will be of any serious interest, at least to medical radiologists.

DISCUSSION ON

"TECHNICAL COLLEGES AND EDUCATION FOR THE ELECTRICAL INDUSTRY"*

Before the NORTH-WESTERN CENTRE *at* MANCHESTER *4th December*, 1951, *the* SOUTHERN CENTRE *at* PORTSMOUTH *7th May*, 1952, *and the* EAST MIDLAND CENTRE *at* LOUGHBOROUGH, *19th January*, 1954.

Dr. P. F. R. Venables (*at Manchester*): The author suggests that technical colleges should not provide engineering degree courses of any type, either full-time or part-time. It is very difficult to find any sound reason for this wholly exclusive recommendation. In Section 4.1 the author states that "Technical colleges are *better* suited for a different type of course," and he also mentions courses for which the technical colleges are *best* suited. The author is, in fact, stating that technical colleges are not better or best suited, but are exclusively or only suited, for courses other than degree ones. Recently in technical education we need to be defended from some of our friends who have been urging that technology should be taken from the universities. This is an absurd plea and is no better founded than the one that the author is making. Both suggestions are all against the evolution of courses in response to needs in this country, and such recommendations have a delusive simplicity.

In Section 4.1 (*d*) the author mentions that many technical colleges providing degree courses have small numbers of students. There are altogether about 220 technical institutions of one kind and another, and of these only 54 are recognized by London University for external degree work. If we assume that the author's arguments are true in part that teaching resources are being dissipated to some extent, and if the number of colleges is reduced by half to say 22, this is a very different matter from recommending that they should all be excluded, except for those cases which the author mentioned where there is a local affiliation. I am interested in this exception, because I believe that even he found it too difficult to ignore.

The author emphasizes the need for degree-course students to have an education of the whole man, then proceeds to make the same argument for all professional engineers, and finally produces the argument with which we can all agree, i.e. that students attending sandwich courses should also have an education of the whole man. It is clear that the author would wish sandwich courses to be conducted in institutions of a sufficient standard of amenities and range of work to make this education of the whole man possible. Having attained these standards within the institutions he then proposes to have the degree courses abolished, when in fact they would enjoy the same facilities and amenities!

There is a very great temptation to use an argument which is based on the best of one side, say the best of the universities, and the worst of the other side, e.g. the worst-equipped of the technical colleges which only just manages to maintain a degree course. Different comparisons could be made. For example, a number of the provincial universities have a far lower percentage of students in residence—about 15% as compared with, say, Loughborough College, which has nearly 80% of its students in residence. I deplore the mutually exclusive attitude which seems to require exclusion of one type of course from another, regardless of the needs which have given rise to them. The same argument would lead to the exclusion of all non-degree work from the universities, and thus impoverish both the universities and society by the loss of valuable diploma courses.

* HASLEGRAVE, H. L.: Paper No. 1219, November, 1951 (see 99, Part I, p. 115).

Mr. N. G. Treloar (*at Manchester*): The author mentions "guidance," and this is most important. It should start in the schools, where headmasters should advise boys as to whether they have a reasonable chance of entering a university, or alternatively of entering industry and obtaining their practical and technical experience in parallel.

Guidance is also necessary for those in industry in order to ensure that they tackle a course of study within their capabilities. A great deal of teaching time and space could be saved if many of those who continue to make unsuccessful attempts on the National Certificate were diverted to the City and Guilds Certificates. Mention has been made of external degree courses. While I support the full-time course, I am very doubtful of the wisdom of the part-time course. It is unreasonable to expect a man to work part-time on a course of studies which take the full-time man four years to cover. The result of continuous evening and week-end study is that insufficient time is left for leisure, and I feel that the more valuable man in industry is one with a Higher National Certificate and a good social development. Furthermore, there are now many avenues to full-time education through scholarships designed to help those whose academic record is up to the required standard, but who, owing to economic difficulties, would not otherwise be able to obtain a degree.

In Section 3.1 reference is made to laboratory work and reports. Technical colleges should pay full attention to this section of a course and not, as has been suggested, leave the experiments to be worked in at the end of the year. In one particular case, no marking of the students' work was done until the end of the session, and marks were lost for unsatisfactory reporting which should have been corrected after the first experiment.

Mr. J. F. Yates (*at Manchester*): The author believes that technical colleges have a special sphere of their own, and appears to believe that the sandwich type of course would be particularly suitable for their engineering students. The award for the satisfactory completion of such a course would be the college diploma and, possibly, the "First Award" of the College of Technologists, suggested by the Report of the National Advisory Council on Education for Industry and Commerce.

This award could not be called a degree, because the universities would not recognize such an award except when made by a university; but the Report makes it quite clear that the award should be designed "at first level (Associate) to provide an educational qualification comparable in value to a university degree, and at the second level (Member) a qualification attainable after an advanced course of post-graduate study and/or research in a phase of technology." The Report was made after consultation with, amongst other people, ten Regional Advisory Councils for Further Education in England and Wales, and it was the opinion of the North-West Regional Committee that the First Award should permit its holder to proceed to a university in order to take a higher degree, and conversely, that the holder of a university first degree should be able to proceed to the second-level award of the major technical college. It was envisaged that there would only be a small number of major technical colleges making these awards. The Report does not

include this prospective interchange, but there is no doubt that the value of the Award must hinge on this possible interchange, and industry and the students themselves will be quick to appraise it on this basis.

With reference to the suggested Higher National Certificate courses for the two kinds of technicians, I feel they should be differently named in order to avoid confusion with the generally understood Higher National Certificate in Electrical Engineering, because recognition of these suggested courses by employers and The Institution is bound to be at a different level from that of the Higher National Certificate, which provides part exemption from The Institution's Examination.

Mr. L. H. A. Carr (*at Manchester*): In Section 3 the author suggests that the future of professional education on a part-time basis is in some jeopardy. This is surely not so, since those most competent to judge appear to appreciate the advantages of the student simultaneously receiving both practical training and scientific education.

An incidental advantage of the National Certificate Courses is that, since many of the part-time teachers are engaged in industry during the day, closer contacts can be maintained between the full-time teaching staff and industry. Such contacts are of importance in keeping full-time teachers up to date. Also, as mentioned in Section 5.1, the value of research from this aspect is frequently greater than the mere worth of the particular results obtained. It is also most necessary that teachers should keep abreast of progress in their particular subjects by reading and study of current scientific literature.

In Section 3.1 the author points out the difficulty of arranging for sufficient tutorial work in the case of part-time students, but this difficulty must be overcome. Not only does the guidance of a tutor prevent much waste of time in exploring blind alleys, but it provides the opportunity for that "reaction of mind upon mind in discussion and debate," which the author in Section 1 rightly stresses as so necessary.

In Section 4.2.3 the author deprecates homework of the orthodox type, but this home study is essential. It is a well-known truism that to understand a problem fully, students must work through it for themselves, and not merely copy into a notebook figures and formulae presented to them on a blackboard.

In Section 4.2 the author puts forward for consideration a "different type of course." Does this not really amount to adverse criticism of existing university and H.N.C. courses?

The author states that one of industry's needs appears to be for engineers possessing a sound knowledge of the fundamentals of science, etc. But surely this is industry's *principal* need, and when in Section 4.2.1 he continues by stating that "the aim should be to give him sound basic knowledge that should last him all his career," is he not merely repeating the avowed intention of those existing full-time and part-time courses that lead to the educational requirements for Associate Membership of The Institution?

Mr. E. Roscoe (*at Manchester*): The author makes the point that students in technical colleges should spend a great deal more of their social life within the influence of the college itself. I strongly deprecate this suggestion. Much has been said about human relations, and one of the problems at present is that our society has gone into strata which confine themselves to their own social contacts with the result that each stratum is a complete stranger to the others.

Mr. T. McGreevy (*at Portsmouth*): In connection with sandwich schemes, the question of "thick versus thin" sandwiches is by no means easy to settle. If students are away from college for a long period, such as 12 months, considerable revision is often necessary when they resume their studies. If the alternate periods are very short, a student never develops that kind of loyalty to either his job or his college which is one of those invisible assets that count for so much in practice. In all sandwich schemes, the question of remuneration during the college period is difficult, particularly towards the end of the course when the student is an adult. I would like to see further inquiry into the two days per week, five-year type of course, and would appreciate particulars of any such courses now being held, with details of the qualification obtained at the end of such a scheme.

Air Comm. W. C. Cooper (*at Loughborough*): I am far from satisfied that the partnership of technical colleges and industry is nationally a really effective one, in spite of those provisions to which the author has referred, i.e. the fact that the majority of technical-college teachers have been trained in industry, that there is a large proportion of part-time teaching provided in technical colleges by people from industry, that there is a regular flow of part-time students from industry through the college, and that many actively engaged in industry serve on advisory committees and as members of governing bodies. I agree that all these provisions can be valuable, but suggest that they are not in themselves enough to secure, let alone maintain, a solid partnership based on factual knowledge and a real appreciation of each other's *modus operandi* and each other's current needs and problems. In these days of rapid industrial development, I stress the word *current*. I feel indeed that, far too often, such reassurances of collaboration are used as a facile excuse by those who pay lip service but little else to the conception of this partnership. Much more is needed than just these. For example, I suggest that increased staffing of technical colleges, revisions of syllabuses and adjustment of time-tabling are essential, not only for the reasons given by the author but also to provide much greater opportunity for all the teachers individually to create and develop their own personal contacts with the industries whose needs they serve; indeed, I would make it a condition of service that these teachers should spend a period of time, at least once every two years, working in, and not just observing, an industry related to those subjects which they teach. Industry, for its part, must be not only willing but really desirous of employing these men effectively, and I think that ultimately it must even go to the extent of releasing, for appreciable periods of continuous time, competent members from among its own employees for service as full-time teachers in technical colleges. In fact, a formal exchange system would have its merits.

Ideally, of course, the author is correct when he states that technical colleges should not provide engineering degree courses of any type. Sometime we may attain that ideal, but I feel that the external degree has proved to be too beneficial and too utilitarian to be thrown overboard easily. However, I am not unduly worried about this problem, since I believe that it will tend to solve itself, if only in an oblique fashion. Apart from differences of opinion on minor points, such as the inclusion of the heat engine and hydraulics in a syllabus intended to meet the needs of my own industry, I like the idea of what the author calls "a different type of course," until he details some of the ways in which he suggests that such a course might be run, i.e. some of his sandwich arrangements.

There seems to be a growing taste for the type of course which alternates somewhat protracted exclusive "study" periods with periods of almost exclusive "application." Some of the reasons given for such an approach are the difficulties of industry in releasing students for two days each week, the difficulties of the day-release student in adjusting his attitude to study, and the more effective development of a corporate spirit among the students.

Taking these three reasons alone: To the first I would answer

that such difficulties as do exist probably arise inherently in those parts of industry which still erroneously regard a student as a productive unit and not primarily as a student. On the second, I would comment that this difficulty can be much exaggerated, and in fact, I would quote from the paper, where it is stated, "It must be remembered that these students are in direct day-to-day contact with industrial applications and are thus better able than full-time students to understand the implications of the principles which they learn in their lectures." To the third, I would simply reply that any industrial organization which fails to use every means within its power to develop a corporate spirit among its own employees is denying itself one, if not the finest, incentive toward industrial efficiency, and any effort on the part of technical colleges towards the same end cannot be more than something additional. I am sincerely willing to be wholly convinced about this particular form of sandwich course, and can understand its value to those who after training will not necessarily be geared directly to the productive effort in industry. However, for those who will be so geared, i.e. the great majority of professional engineers and technicians at least in the manufacturing industries, I must underline and support the author when he states: "The closest possible linking of academic work with industrial training is necessary as well as careful and original planning of both technical and practical training." I cannot yet see that periods of absence for one month, one term, or six months at a time, from the production atmosphere and tempo can be regarded as "the closest possible linking" or likely to produce men who are as productivity-minded as those who have been in day-to-day contact with industry throughout all the years of their training.

I think the author has omitted one most important and significant educational service that technical colleges can provide for industry. There is a growing need at present, at least in the manufacturing industries, for semi-skilled operatives, i.e. men and women who have not been trained as craftsmen, to learn more than they have done in the past—which has very often been nothing—of what lies behind the operations they perform. I have in mind particularly an experiment which is taking place in a neighbouring town, with semi-skilled operatives engaged in the manufacture of plastic mouldings and in the application of anti-corrosive finishes such as electroplating and enamelling to wood and metal. A technical college is co-operating with an industry to give these people some knowledge and appreciation of the principles that underlie the processes. The industry has already found an improvement in the work of the operatives, a more intelligent approach both to the requirements and the difficulties that arise, and a lessening in the burden of supervision; all very desirable contributions not only to productivity but to the self-respect of the employees concerned.

Group Capt. E. J. Bradbury (*at Loughborough*): I am very pleased that the primary emphasis in the paper is on the necessity for any system of education to develop the whole man. The detail of individual courses can be laid down to meet individual requirements, but we must not lose sight of the absolute necessity to produce from our courses a balanced person.

When talking to our student apprentices, I sometimes think that we tend to forget what hard work is involved in the attainment of the National Certificate. The Certificate represents the be-all and end-all of their life, and the fear that they may not get it is always present. To suggest to these young men that they should widen the scope of their studies to include subjects other than those necessary technical ones would seem to them to be quite mad, and yet it has got to be done and steps must be taken to ensure the development of the whole man as well as a competent engineering technician.

The technical requirements of the National Certificate are extensive—so much so that I have the suspicion that we might get better results with fewer subjects or possibly a lower level in the examination. Whatever the solution, we must not lose sight of that contact on which the author laid primary emphasis, and we must develop a balanced personality of competent technical skill if we are to meet national requirements from the engineering industry again.

Prof. J. A. Pope (*at Loughborough*): The author states that he is not in favour of technical colleges preparing people to study for university degrees. I think he is right, but in spite of generous grants, there is a group of people who cannot afford to go to the university, since there is a means test for those who can obtain a university grant. This hits those with large families, and they must not be overlooked.

The author has stated that the function of the technical college is peculiarly its own. When I look at his course, he has exactly the same total number of hours of instruction as we have for the university full-time degree course. I imagine that the qualities of the entrants he will want for his course will be much the same as those the university is asking, although the training may be somewhat different. Therefore, I would suggest that changing the title does not solve the problem.

We must persuade schoolmasters that an industrial career is not only a good career but is of vital concern to our future existence. Free secondary-school education is now available to boys who could not previously have had this type of education, but the parents of these boys have generally had no experience in planning a child's future. They are entirely in the hands of the schoolmaster, and it is significant that only some 10% of all university entrants are trained as technologists. This is one of our big problems.

The problem of research is extremely difficult to solve. It depends not only on having people of ability, but also on creating an atmosphere in the institution itself. The technical colleges have not yet got this atmosphere, and I do not know how they are going to get it. If senior technical-college institutes were set up, it might be possible to create the right atmosphere. It is said that universities should do fundamental research and technical colleges should do applied research, but if you are doing applied research, that in itself will bring to the surface fundamental problems which will have to be solved. It is impossible to separate the two branches like that.

Prof. H. Cotton (*at Loughborough*): Education is indivisible. The young man who comes to the university is not educational raw material. He has been processed at school for three-quarters of the total time during which he is educated.

When I interview young men who come to the university, I ask them many questions, three of which are:

How good are you at mathematics?
What kind of mathematics did you do at school and what kind of science?
What kind of advice did your headmaster or teacher give you?

The answer to the first question is generally that he has done pure mathematics and not applied mathematics. He is handicapped because he has not done this.

Secondly, he has not done physics to a high standard, but he has done general science. This is inadequate for a young man who wants to be a scientist or engineer.

Thirdly, the headmaster gave no advice about how to be an electrical engineer. Headmasters often have no interest in engineering; they regard it not as a profession but a trade.

If you compare a set of old examination papers with present-day ones, they are elementary, and thus the burden on the student becomes progressively greater. Education should be an evolution, but it should be a slow evolution so that the student can keep pace with the development of his subjects.

I think the term "external degree" is somewhat unsatisfactory. I suggest that the diplomas awarded would have a higher status if there were external examiners. If a man has a degree it is something which everyone can assess. If it is a diploma people "look down their noses at it." Why does a man want a degree? Is it for snob value or is it a very valuable asset in industry?

Dr. H. Buckingham (*at Loughborough*): In Section 8.2 a change of outlook in technical colleges is demanded. I believe that certain changes have been taking place, even since the paper was first published. One indication of this is a certain change in attitude towards the provision of university degree courses. From the majority of technical colleges a degree can only be obtained under the external examination system. In covering a large number of approved colleges the system inevitably involves some lack of flexibility in the content of syllabuses and a measure of remoteness of control which has no counterpart in an internal scheme. Moreover, the need for students to submit a large number of formal laboratory reports imposes an added burden both on students and staff. Except in a very few colleges, where a widespread demand for such courses still exists, there seems good reason for the author's contention that degree courses should not be attempted.

However, it is important to remember that one outstanding benefit of the external degree system is its influence on teaching standards. A high level of instruction must be maintained if the exacting examination requirements are to be met. It follows that any scheme for an alternative qualification can be successful only if it is backed by a means of preserving the standard of the work. Furthermore, the measures adopted for the purpose must readily be acceptable outside, as well as inside, the technical colleges. It is probable that a careful application of the author's suggestion of appointing external examiners would meet the case, provided that a ready interchange of views were made possible between the examiner and the college staff.

Mr. J. C. M. Sanders (*at Loughborough*): The need to build character into our young men, to introduce them to the humanities, and to give them some insight into commerce and administration does not appear to be in question, but the difficulty is how to do it when they are already being crammed with technical education. I suggest that the time is ripe to investigate the possibility of National Service training being modified to meet our requirements, and at the very minimum to ensure that there is no abrupt cessation of the educational programme.

Surely the best service a young technician can render the nation is to fit himself adequately for his career. In far too many cases the present form of National Service does not help him to do this. Could not The Institution Education and Training Committee investigate the whole question of further education during National Service in order to ensure that it rounds off technical training, develops character and ability, and returns the fully qualified man to industry?

Mr. R. C. Woods (*at Loughborough*): The relative standing of the degree and the diploma has been mentioned. If we neglect the snob value, in industrial recruitment we look to the present reputation of the school of a university or the department of the college from which the applicant comes, in relation to the appointment to be made. A technical college which turns out men of quality need not worry about the status of its diploma.

I strongly support the plea for teaching staff to specialize, if they wish, in grades of work. Some teachers, who are very able in the higher branches, have no sympathy with the more elementary stages and little or no appreciation of the difficulties of the junior student. They are much more interested in things than people. Excellent as they may be to lead the more developed mind, to the less mentally agile they are a menace, for the students are left without the one essential to confident work—a grasp of first principles. A much less erudite man, with a knowledge going little beyond the stage he teaches, but awake to his pupils' needs and prepared to try more than one mode of approach to get an idea accepted, is much more useful.

Mr. C. A. Brearley (*at Loughborough*): I should like to endorse the author's plea in Section 2.1 for a pruning of syllabuses. There is a tendency in some quarters to pack syllabuses in such a way as to demand from a student, for examination purposes, a mass of factual knowledge over a wide field, to the detriment of the development of habits of thought and judgment. The aim, as suggested in the chart, of age groups in which, in general, professional academic qualification should be completed by about the age of 22 years, is desirable. In the early years of his adult life an engineer, like other citizens, should be free to devote his leisure to advancing his knowledge in his own sphere of work, clear of the incubus of examinations, and to participate in the domestic, civic and religious obligations and the pleasures of his environment.

Referring to Section 5.1, most of us need only recall inspiring teachers of our schooldays to counter the statement that research is essential to "live" teaching. It is more important for a teacher to try to keep abreast of engineering progress over as wide a range as possible than to undertake research in a limited direction; it is an advantage, of course, if the time and facilities for research are available.

Dr. D. A. Jones (*at Loughborough*): In the paper it is inferred that the university or technical college is not greatly concerned over the student being given a sound practical training. If the college took an interest in the works training, would not a certain continuity of education be obtained? After all, apprenticeship is part of the education of an engineer.

The scheme could possibly be effected by the industrial concern sending a quarterly report to the university or college on the progress of the student. In return, the college could reply with comments derived from experience of the student over the previous three or four years.

Dr. Gibbs *et al.** stated that the Higher National Certificate student takes longer than a graduate to develop into a useful engineer. There are obviously many possible reasons for this slower development, but it does seem that greater emphasis could be placed on the facilities for the social and athletic activities of the H.N.C. student. In this way he acquires the ability to communicate his newly gained knowledge and to absorb new ideas by social intercourse.

Mr. E. Houghton (*at Loughborough*): The yearly intake into a technical college embraces a wide field, from the man running his own radio shop, through the large manufacturers and nationalized industries, to every non-electrical industry which employs electrical staff. The college prospectus offers a range of courses calculated to meet this diverse demand with economic classes. Many syllabuses of national examinations have been framed by trade advisory committees, who are often more concerned with what ought to be included than with what can be assimilated in a limited part-time attendance. Anybody who has attended the enrolment at a large college will know how little time is available for individual attention to each student. The teaching weeks before the examinations are limited, and any attempt at grading in the early part of the course, with consequent upheaval in the days or evenings chosen for attendance, is fraught with difficulties.

Many embark on the wrong course, and many fall by the wayside. At present youths are in short supply, and an increase in the number of State scholarships has still further restricted

* GIBBS, W. J., EDMUNDSON, D., DIMMICK, R. G. A., and LUCAS, G. S. C.: "Post-Graduate Activities in Electrical Engineering," *Proceedings I.E.E.*, Paper No. 1265, February, 1952 (99, Part I, p. 161).

the numbers available for apprenticeships. Industry cannot afford this wastage.

I ask for two things. First, from those concerned, either through trade association or by academic position, with framing courses, let us remember that every school leaver is not from the fifth form of a grammar school in the pre General Certificate of Education era, and that there are late developers who will be able to move from one course to another. Let craft, technician and professional courses dovetail one into another, so that a person with outstanding results in a technicians' course does not have to start right at the bottom of a professional course. Secondly, I ask employers and training officers concerned with apprentice selection to include in their selection tests an attainment test, which will ensure that the selected candidates can benefit from a chosen technical course. I would be happy if, like the university matriculation, there were agreed school-leaving standards, although as a citizen I would object to any formal labelling. Any technical college would, however, be only too glad to supply suitable sample papers to employers, and would welcome any pre-testing and block enrolment. These considerations apply to apprenticeships designed to reach H.N.C. level in, say, four or five years, with exemption from the early years of the course on account of the G.C.E. (which is still a very variable standard), and they also apply to specialized craft courses where the student's early interest is obtained by specialist practical instruction and where a general pre-senior course would stifle interest and progress. Normal entry to National Certificate courses through suitable pre-senior courses does afford an opportunity for selection. However, care must be taken with these, lest through absence of laboratory work or guidance from engineering teachers the adolescent falls away, because the essential connection between course and work is missing.

Finally, I would reiterate the great need for all members of The Institution to keep in touch with schools, school-leavers and parents to advertise the opportunities in engineering and to ensure that, in competition with all careers, we in the electrical industry get our fair share of the best possible boys.

Messrs. R. F. Marshall, F. W. Taylor and **F. G. Copland** also contributed to the discussion at Manchester, and **Messrs. O. S. Woods, G. E. Smith** and **B. Gill** also contributed to the discussion at Loughborough.

Dr. H. L. Haslegrave (*in reply*): Whilst it was emphasized in the discussions that the need for the interchange of staffs of technical colleges and industry is even greater now than it was some years ago, it must be realized that the difficulties of effecting this interchange have also been increased by the rapid industrial development. It is not likely, for instance, that a teacher can go into any works and immediately justify the payment of a salary equivalent to his teaching salary. Neither can an industrial executive settle down immediately to full-time teaching work. Three ways of making the interchange of staff possible and valuable are worth mentioning.

The practice exists in a few colleges of allowing some staff to spend a half day, or more, each week in a works, at first observing and later actually participating in the work. Current developments and current problems are thus effectively brought into the college, and the transition to the teacher going on to the pay roll of the works for an extended period is facilitated. For some sections of the teaching work it is better that a teacher should leave his college and go for an appreciable time—two years or more—into industry. It is now possible for pension rights to be safeguarded in such cases, but the teacher concerned would normally desire to return to a higher-status post in his original, or another, college.

In a few colleges teachers are allowed to carry out consultative or investigating work for industry, which keeps them in close touch with industrial conditions and so makes transference to full-time work in industry comparatively easily possible. This practice could well be extended.

The third way, operating for at least one university, is the scheme of a group of students spending some weeks in an industrial concern, and carrying out some project under the guidance and with the aid of members of the university staff. Such staff acquire a good and close relationship with the concern.

One cannot leave this problem of ensuring that teaching staff have a good knowledge, as well as experience, of industrial conditions, without mentioning that the tendency of companies to compete with each other for the recruitment of graduates by offering increasingly higher salaries for post-graduate apprenticeships is making it more and more difficult for colleges to obtain suitable staff. Colleges often require a man to have had at least two years' training and experience in industry after graduation from a full-time course or sandwich course before taking up an assistant lecturership, and yet that man may have received during his practical training a salary equivalent to, or greater than, the maximum salary he would receive as an assistant lecturer after twelve years of teaching. If such men can only be recruited at the lectureship grade, the standard of teaching at both assistant and lectureship grades will fall. Thus industry, by increasing the financial incentives it is offering to graduates, is making it very difficult for colleges to supply the required type of man.

Several speakers stressed that colleges should develop men with balanced personalities, but perhaps those who guide industry and the professional engineering bodies do not appreciate fully that, as they increase the technical attainments required of students on completion of their college courses, they make it increasingly difficult for time to be given to the development of personalities. The plea made in the paper, and justifying repetition, is that, after taking into account National Service requirements, there is a definite limit to the time that can be spent on training, and a suitable proportion of this time should be given to the development of personality. It is possible for the student to attain only a certain standard of technical ability and knowledge by using the remainder of the time, and if industry or the professional engineering bodies require a higher standard, it should be secured by means of post-graduate study.

In the comments made during the discussions upon the proposals contained in the paper for sandwich courses, there was evident some misunderstanding of three of the essential points in such courses:

(*a*) By incorporating in the courses extended periods of attendance at the college (longer than two days per week), full corporate college life is made possible, and becomes an integral part of the course.

(*b*) Close correlation of principles and practice becomes possible through partnership of teaching staff and the personnel in industry responsible for the training of the apprentice, in both the planning and the operating of the works training, and through the teachers contacting the apprentices in the works.

(*c*) Production consciousness is retained throughout the complete period of training.

The title and status of the award was naturally referred to frequently, but this problem appears nearer solution now than at the time of the discussions.

Recruitment of students was touched upon by many speakers, in two of its aspects—recruitment to the individually appropriate

type of course, and recruitment to the engineering industry. The first aspect involves not only the planning of the various courses, but the giving of the best advice to students and, in their early years, their parents, by college staff and industrial staff, and it also requires the readiness of education authorities to make financial grants to students taking sandwich courses. Some authorities have become more generous in this last respect during the two years 1953–54, but many still award grants only for degree courses. Whilst a main purpose of sandwich courses is to provide a fuller training than through the medium of part-time day courses, they are also likely to attract into them, and thus to the engineering industry, students who would not have considered going to a university and who might have sought another career rather than take part-time engineering courses. There is no single method of ensuring increased entries to the engineering industry, and all who are called upon to advise and guide youths should be made aware of the purpose of the industry and the possibilities in it. Headmasters of schools of all types, careers masters, parent-teacher associations and parents are all vital in this matter.

Individual members of staffs of technical colleges and industrial concerns also play an important part through their separate personal contacts. It should be appreciated that many people remember that only 25 years ago first-class engineers, technicians, craftsmen and apprentices were discarded, when trade was bad, by organizations which at present are operating elaborate schemes for the recruitment and training of apprentices of all types. Some of those who were so discarded are now parents of potential apprentices, or are called upon to advise potential apprentices. Only time can remove completely this shadow of the past that is still affecting recruitment to the industry.

DIGESTS OF INSTITUTION MONOGRAPHS

THE RESIDUAL TIME-CONSTANT OF SELF-SATURATING (AUTO-EXCITED) TRANSDUCTORS

621.318.435.3.011.6 Monograph No. 140 M

ULRIK KRABBE, Ph.D.

(DIGEST *of a paper published in July*, 1955, *as an* INSTITUTION MONOGRAPH *and to be republished in Part C of the* PROCEEDINGS.)

The self-saturated transductor has a close electric analogy to a d.c. shunt machine. When a direct voltage is applied to the field winding of a d.c. machine, the output varies according to the equation

$$V_N = k_v V_s (1 - \varepsilon^{-t/\tau})$$

Here V_N is the output voltage, V_s is the voltage applied to the field winding, and τ the time-constant. It is well known that τ includes a part proportionate to L/R for the field winding but also that it depends on L/R values for other closed circuits around the poles (and also on eddy currents in the pole iron).

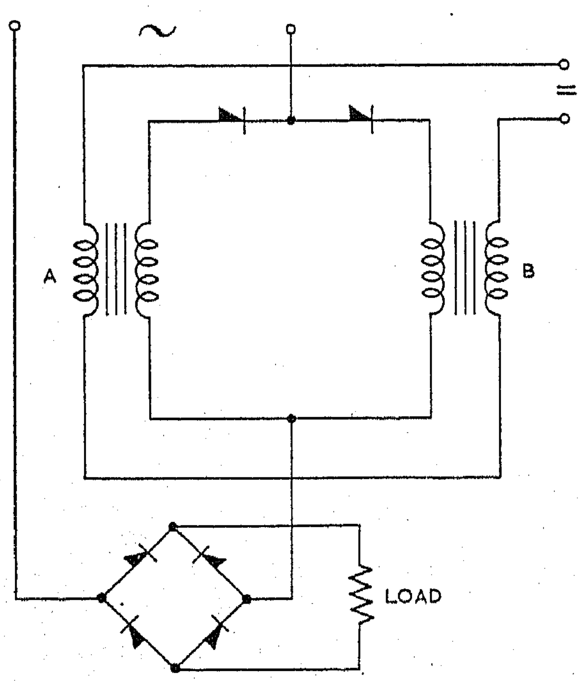


Fig. 1

Consider now the self-saturated transductor circuit, Fig. 1. For this coupling we have very similar laws for the time properties, but the case is more complex because the power winding itself forms a closed circuit, so that it contributes to the time-constant of the transductor. However, there are two rectifiers in the closed circuit, and the influence of the power winding on the time-constant depends on the lengths of the intervals during which both rectifiers are conducting, because when there is a blocking voltage across one of the rectifiers no circulating current can flow and there is no contribution to the time delay from the power winding. When the duration of reverse voltage across the rectifiers in each half-cycle is known, the overall time-constant of the transductor can be calculated.

A qualitative analysis shows that the duration of the reverse voltage depends on the time-constant of the control winding. A very complicated variation dependent on non-ideal properties of the cores is to be expected, and measurement has shown that the cut-off angle is quite close to 180°, so that it is only during a small fraction of each period that both rectifiers are conducting.

The investigation showed good agreement between the measured time-constant of a transductor and the time-constant calculated on the basis of measured reverse-voltage intervals. The time-constants were measured by the frequency-response method with small signals and measurement of phase displacement between the signal voltage and the signal-frequency component of the output voltage. The time-constant was defined as the inverse value of that signal angular frequency which gave 45° phase displacement.

It proved that only very small signals could be used because the reverse-voltage interval for the rectifiers is easily influenced by a signal. The result is that a signal which increases the output tends to decrease the time-constant, and a signal which decreases the output increases the time-constant.

The measurements were made with conditions as ideal as possible, using hot-cathode vacuum tubes. The non-linear characteristic of selenium rectifiers is assumed to influence the time-constant, because the relatively high resistance at low forward current reduces the time-constant of the power circuit.

The circuit of Fig. 2 is more favourable than that shown in Fig. 1, because the reverse voltage across the rectifiers is larger

Dr. Krabbe is with A/B Elektromekano, Hälsingborg, Sweden.

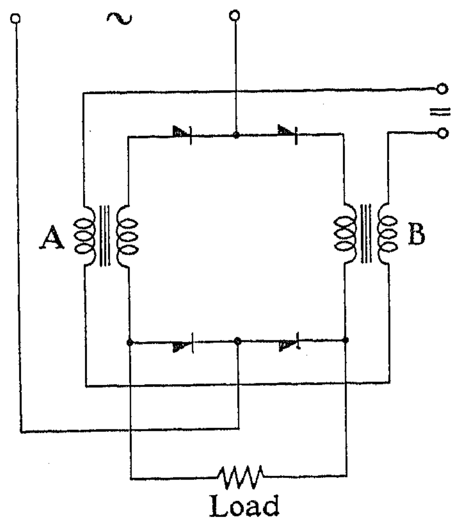


Fig. 2.

and therefore the reverse-voltage intervals are less influenced by large transient signals; also the larger number of rectifier plates reduces the time-constant of the main winding.

The transductor could be made fast-acting at the sacrifice of amplification in two different ways: by increasing the resistance in the control winding or by means of negative feedback from the output voltage. The delay contributed by the main winding will give different action, but in both cases it will increase the time-constant.

BIBLIOGRAPHY

(1) KRABBE, U.: "The Transductor Amplifier" (Munksgaard, Copenhagen, 1949).

(2) MILNES, A. G., and LAW, T. S.: "Auto-Self-Excited Transductors and Push-Pull Circuit Theory," *Proceedings I.E.E.*, Paper No. 1599 M, April, 1954 (**101**, Part II, p. 643).

(3) LAMM, U.: "The Transductor" (Esselte Aktiebolag, Stockholm, 1943).

(4) SCORGIE, D. G.: "Fast Response with Magnetic Amplifiers," *Transactions of the American I.E.E.*, 1953, **72**, Part I, p. 741.

(5) RAMEY, R. A.: "On the Mechanics of Magnetic Amplifier Operation," *ibid.*, 1951, **70**, Part II, p. 1214.

(6) DUNNEGAN, T., and HARNDEN, J. D.: "The Cyclic Integrator. A Device for Measuring the Frequency Response of Magnetic Amplifiers," *ibid.*, 1954, **73**, Part I, p. 358.

FUNCTION GENERATORS BASED ON LINEAR INTERPOLATION WITH APPLICATIONS TO ANALOGUE COMPUTING

621.389 : 517.94 : 518.5 Monograph No. 137 M

E. G. C. BURT, B.Sc., Associate Member, and O. H. LANGE, Dr. rer. nat.

(*Digest of a paper published in June, 1955, as an* INSTITUTION MONOGRAPH *and to be republished in Part C of the* PROCEEDINGS.)

The electronic type of analogue computer is now used extensively as an aid to the solution of a variety of problems. The computer in its simplest form consists of an assembly of d.c. amplifiers, which may be operated as integrators or as summing units. Such a machine can be used to solve linear differential equations with constant coefficients, of order equal to the number of integrators available: this, however, is of limited use, since the solution can readily be obtained analytically.

The scope of the machine can be greatly extended by the use of function generators, i.e. units in which the output voltage is a specified function of the input voltage or voltages; the function may be known analytically or simply as tabulated data. Such generators are essential for the solution of non-linear and variable-coefficient equations, for co-ordinate transformations, and for introducing coefficients which have been derived experimentally.

The function generators described in the paper are based on linear interpolation, in which the linear intervals are provided by diode circuits. The latter are so arranged that the current flowing is proportional to the input voltage throughout a given interval, the current/voltage slope corresponding to the slope of the straight-line approximation to the function for that interval.

The basic circuit is shown in Fig. 1. If the internal gain of the amplifier is very high, the input and output voltages v, v_0 are related as follows:

$$\left.\begin{aligned} v_0 &= -iR = -\frac{(vG_1 - V_BG_2)G_{3D}R}{G_1+G_2+G_{3D}}, & v > V_BG_2/G_1 \\ &= 0 & v \leqslant V_BG_2/G_1 \end{aligned}\right\}$$

where $-V_B$ is a fixed negative voltage, G_1, G_2 are the conductances of R_1, R_2, and G_{3D} is the combined conductance of the forward diode resistance and the resistor R_3.

Because of the high internal gain of the amplifier, the point P (Fig. 1) is effectively at earth potential, and it follows that the

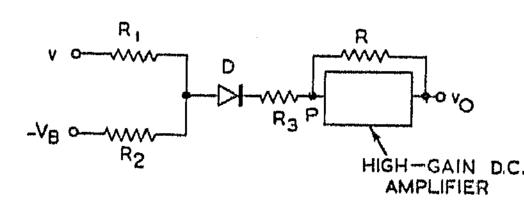


Fig. 1.—Conversion to voltage output.

outputs of further diodes may be commoned at this point. The contribution from each diode will be unaffected by the currents from the remaining units, and the output voltage will be proportional to the sum of the currents.

Fig. 2 shows the arrangement for generating a monotonically increasing function, such as the parabola $v_0 = cv^2$. The number of diodes in group 1 is equal to the number of straight lines used to approximate the curve, and the choice of resistance values governs the slope and cut-off point of each diode circuit. A second group of diodes, connected as shown, is required to deal with negative values of the argument v.

The squaring circuits of Fig. 2 can be used to provide a multiplier, based on the identity

$$4cv_xv_y = c[(v_x + v_y)^2 - (v_x - v_y)^2]$$

where v_x, v_y are voltages proportional to x, y, the variables whose product is required, and c is a dimensional scale factor. The multiplier (Fig. 3) can be so arranged that only three amplifiers are required to give the correctly signed product v_xv_y when v_x, v_y are unrestricted in sign. Either the positive or negative product can be selected by means of a switch, so that the multiplier includes its own reversing element—a useful facility in an analogue machine. Either output is obtained at a low impedance, and, by using drift-corrected d.c. amplifiers with a large voltage swing (± 50 volts), high accuracies can be achieved ($0{\cdot}2\%$ of the actual product above half-scale, and $0{\cdot}1\%$ of the maximum product below half-scale).

Mr. Burt and Dr. Lange are at the Royal Aircraft Establishment, Farnborough.

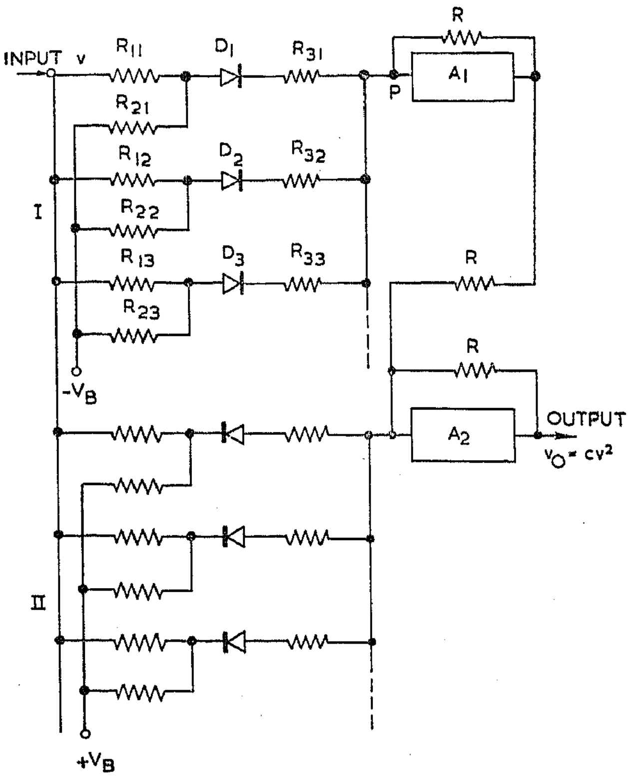


Fig. 2.—Circuit for the parabola.

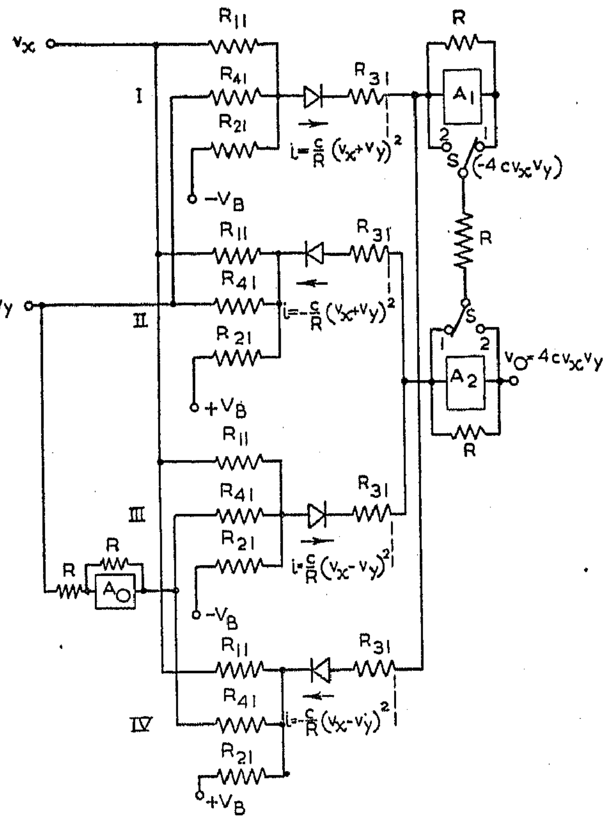


Fig. 3.—Diode multiplier.

The diode technique can be extended to deal with non-monotonic functions, by suitably rearranging the signs of v and V_B, and by reversing the diode connections (Fig. 1). A combination of such circuits provides the means for approximating to a general function, with a positive or negative argument. The paper gives a number of examples, one of which is illustrated in Fig. 4, which shows a generator for the sine function

$$f(v) = V \sin cv$$

over the range -2π to 2π. Each of the groups I–IV contains sufficient diodes to give a good approximation to the sign curve over the stated range, which can be extended by adding further groups of diodes, although at some sacrifice in accuracy.

Fig. 5 shows the results obtained with an experimental unit constructed according to Fig. 4. Seven tangents were used to approximate the sine curve over the range 0 to π, requiring three double diodes [the first tangent is provided by a direct input,

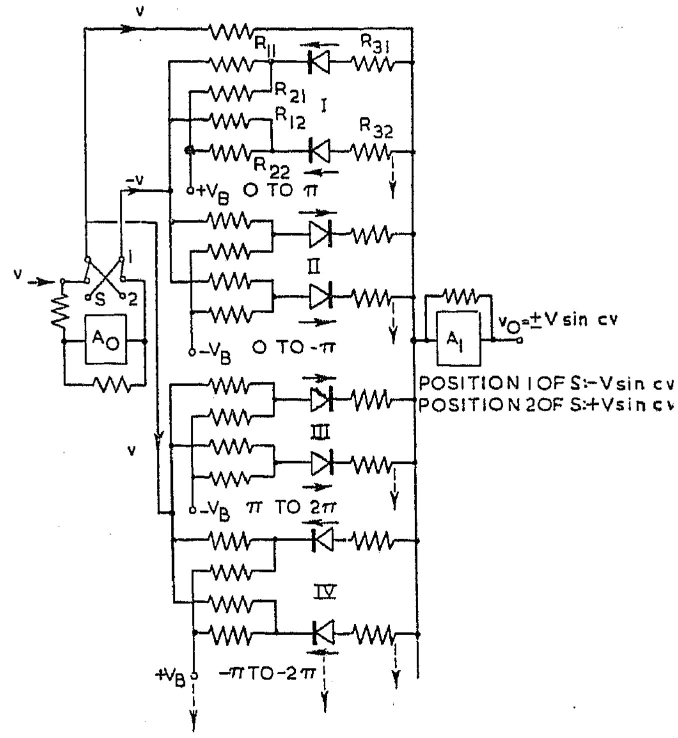


Fig. 4.—Generator for $f(v) = \pm V \sin cv$.

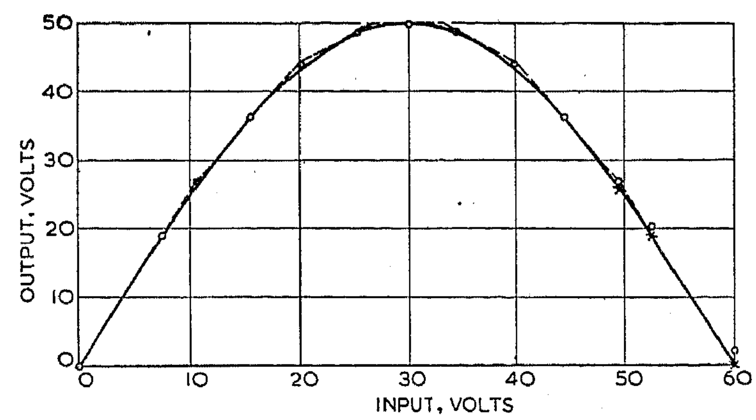


Fig. 5.—Experimental results for the sine function.
⊙ Experimental points with resistors as calculated.
× Experimental points after adjustment of R_{35} and R_{36}.

v (Fig. 4)]. Thus the complete generator contains six double diodes, covering the range $-\pi$ to π. After adjustment of two resistance values, the maximum error from the tangents (Fig. 5) was about 0·5%, and the error from the sine curve about 1·5%. The accuracy can be increased, within limits, by taking more tangents.

The diode function generator has a number of advantages—accuracy, simplicity, flexibility, and a high speed of response. Some care, however, is needed in the selection and ageing of diodes, while the resistors and the bias voltages must be very stable to produce the best results.

References

(1) Lange, O. H., Burt, E. G. C., and Holbourn, C. R.: "A Drift-Compensated D.C. High-gain Amplifier for Summation and Integration," *R.A.E. Technical Note No. G.W.*75.

(2) Harder, E. L., and Carleton, J. T.: "New Techniques on the Anacom Electric Analogue Computer," *Transactions of the American I.E.E.*, 1950, **69**, p. 547 *et seq.*

(3) Marshall, B. O., Jr.: "An Analogue Multiplier," *Nature*, 1951, **167**, p. 29.

(4) Thomas, W. R., and Squires, M.: "Electronic Analogue Methods of Multiplication," *R.A.E. Technical Note No. G.W.*53, *Aeronautical Research Council*, **13**, p. 170.

(5) Lange, O. H., and Herring, G. J.: "Some Electronic Multipliers based on Diode Function Shapers," *R.A.E. Technical Note No. G.W.*245.

A REFLEX KLYSTRON OSCILLATOR FOR THE 8–9 MM BAND

621.385.1.029.6 : 621.373.423 Monograph No. 143 R

D. J. WOOTTON, B.Sc., Associate Member, and A. F. PEARCE, Ph.D., F.Inst.P.

(DIGEST *of a paper published in August*, 1955, *as an* INSTITUTION MONOGRAPH *and to be republished in Part C of the* PROCEEDINGS.)

The paper describes a reflex klystron oscillator, tunable over the wavelength range of 8–9 mm, which is suitable both for use in a superheterodyne receiver and as a source for laboratory measurements. The type number is VX5023.

Mr. Wootton and Dr. Pearce are at the E.M.I. Research Laboratories, Ltd.

As a basis for the design, it was decided to use a cavity having

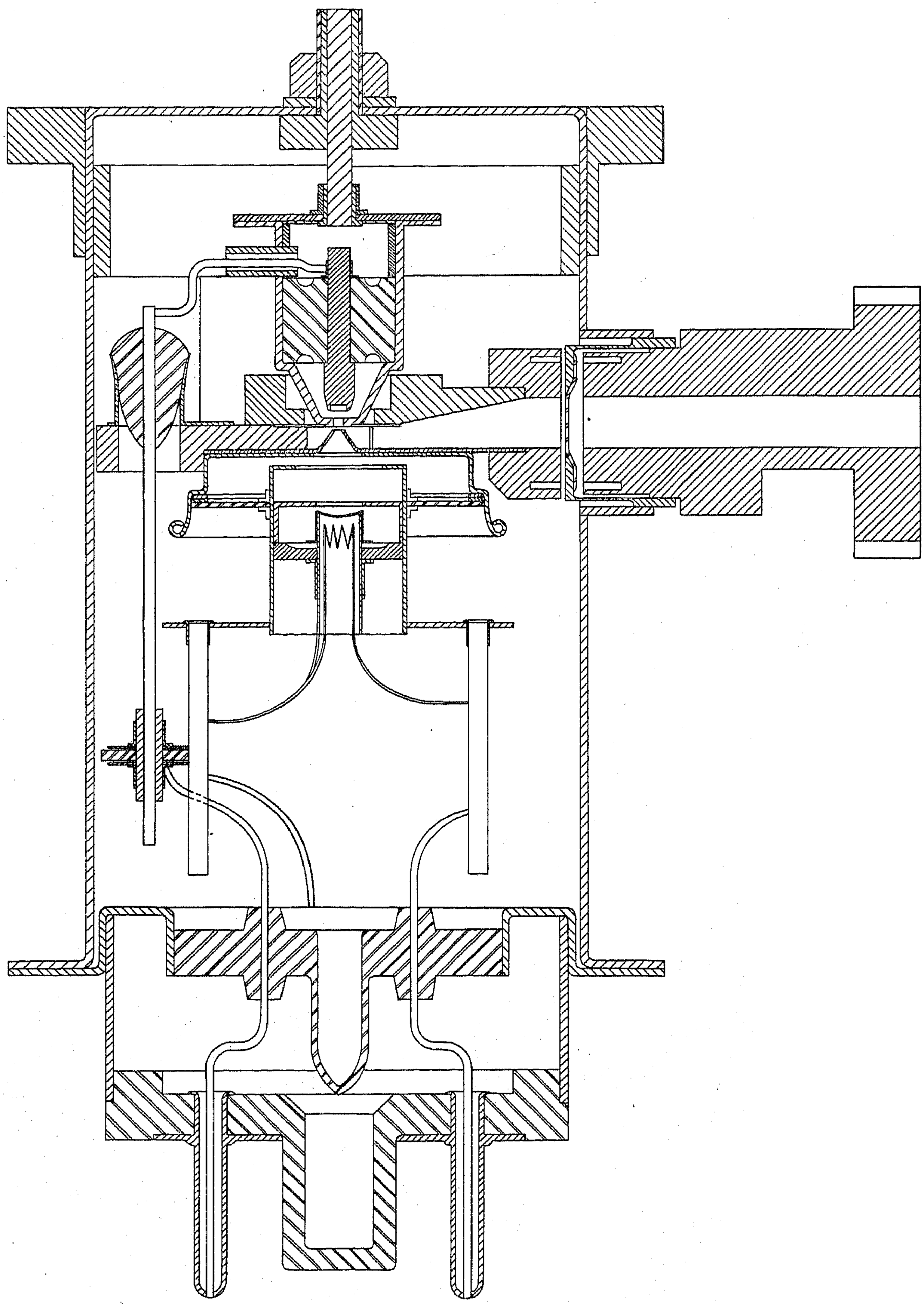


Fig. 1.—Sectional view of valve without tuner.

gridless apertures and operating at about 2 kV. This avoids the difficulty of making a grid which for this wavelength would be very small and might well have been capable of dissipating only a small amount of power, thus limiting the input power and efficiency. The cavity, which is about 4 mm in diameter and 1 mm deep, operates in its fundamental mode, thereby reducing copper losses to a minimum and ensuring a smooth variation of power output over a 10% band. Tuning is accomplished by flexing a thin diaphragm forming one wall of the resonator, thus varying the capacitance at the gap.

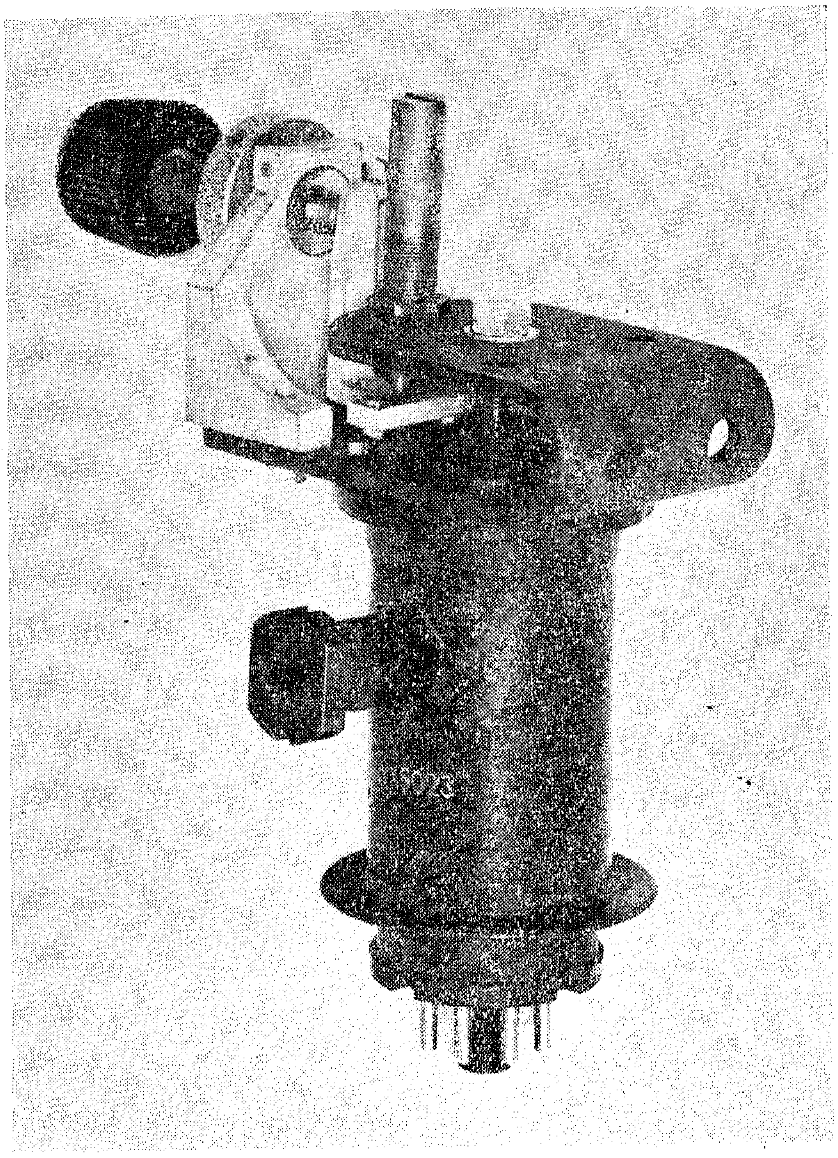


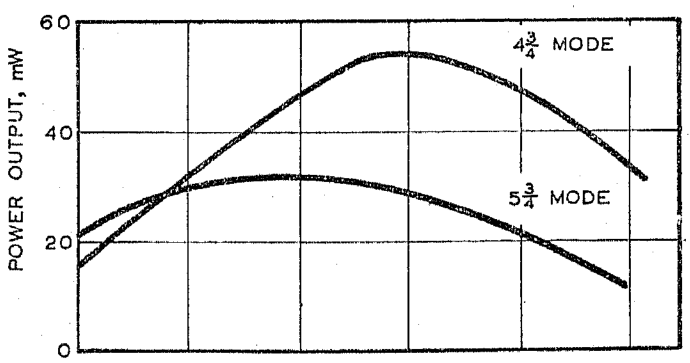
Fig. 2.—Photograph of valve type VX.5023.

The method of arriving at the final design was to determine as many of the relevant dimensions as possible by a combination of calculation, based on the theory of Barford and Bowman-Manifold,[2] and previous experimental knowledge, and to find the remainder by empirical means. The theory indicated that for a minimum power output of 15 mW, and allowing for a reasonable safety margin, a resonator current of 10 mA was required, and for this current a resonator aperture of 0·020 in diameter was chosen. The gun and reflector designs, cross-sections of which may be seen in Fig. 1, were arrived at empirically, the gun being a modification of one used on an earlier valve.[4]

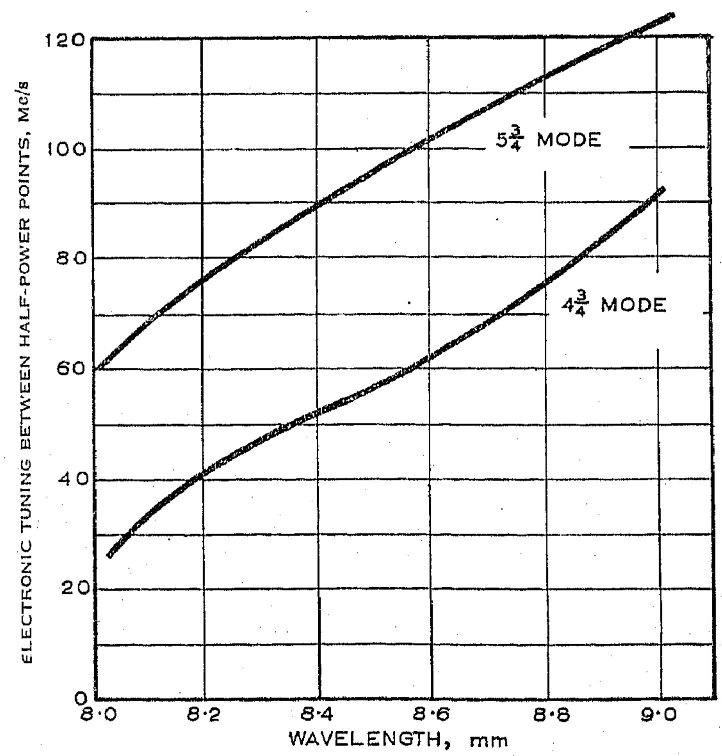
The main requirement demanded of the method of assembly is very accurate alignment of the resonator, reflector and gun, including freedom from tilt. This was achieved by machining the resonator from a copper block, thus providing good locating surfaces for the electrodes. By mounting the resonator assembly inside a thin steel casing, which constituted the vacuum envelope, an advantageous layout was obtained; in particular, the output power could be extracted through a thin glass window sealed to the envelope, into an external waveguide. Moreover, the good thermal contact between the copper resonator and the metal envelope ensured adequate cooling and minimized frequency drift.

For mechanical tuning, a screw-and-lever mechanism was used in combination with a C-spring. This provided a large velocity ratio and ensured a smooth control of the cavity frequency, which is very sensitive to the setting of the resonator gap; a change of 0·001 in gives rise to a frequency shift of 500 Mc/s. The tuner is mounted on top of the valve as shown in Fig. 2.

The variation of power output and electronic tuning range with wavelength is given in Fig. 3 for an input of 2 000 volts, 10 mA. The curves refer to an average valve (maximum powers of up to 160 mW have been obtained with some valves), where the output slot coupling the cavity to the waveguide is 0·060 in wide. This is not, however, the slot width for maximum output,



(a)



(b)

Fig. 3.—Characteristics.

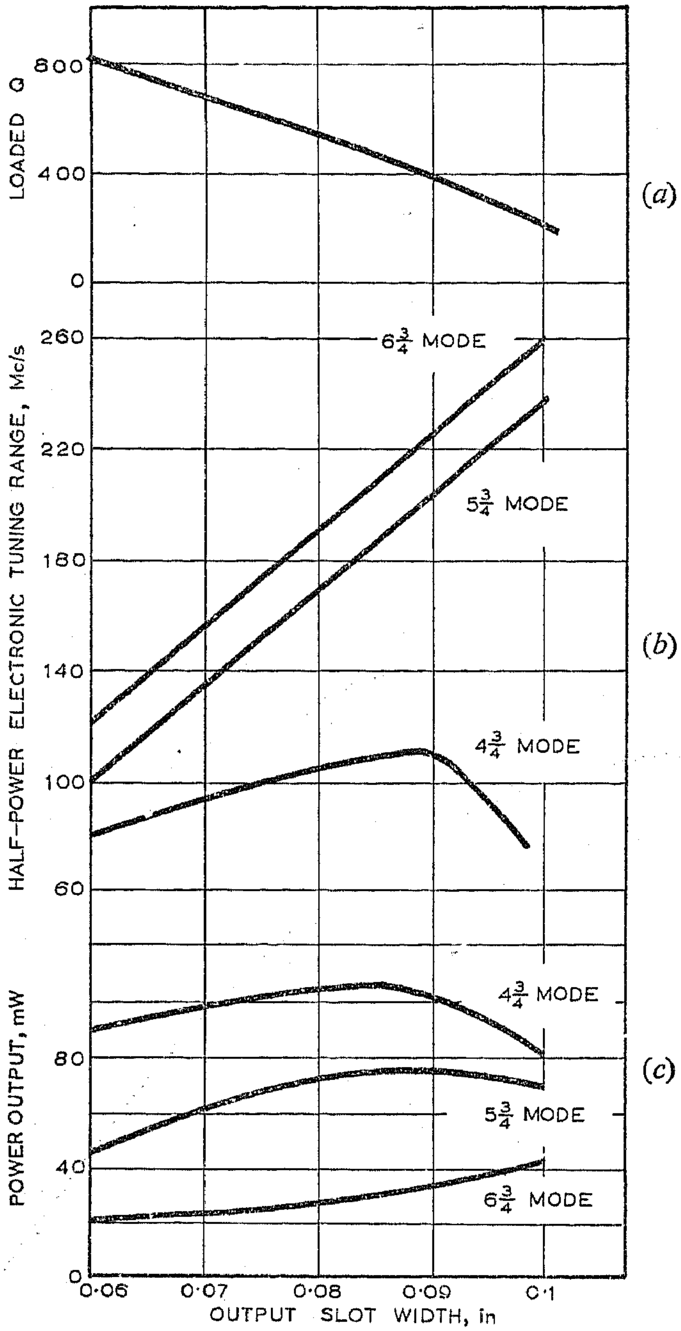


Fig. 4.—Variation of characteristics with cavity loading.

as is shown by the curves of Fig. 4; maximum power and electronic tuning range occur with a slot width of about 0·080 in. A disadvantage of the wider slot, which is discussed in the paper, is that the noise power output from the valve increases with slot width (or with decreasing Q_L), and this consideration is important in mixers operating at this frequency. Slot widths of 0·060 in and, more recently, 0·070 in have been used in manufacture.

References

(1) VAUGHAN, J. R. M.: "A Millimetre-Wave Magnetron," *Proceedings I.E.E.*, Monograph No. 142 R, August, 1955 (**103** C, p. 100).

(2) BARFORD, N. C., and BOWMAN-MANIFOLD, M.: "Elementary Theory of Velocity-Modulation Oscillators," *Journal I.E.E.*, 1947, **94**, Part III, p. 302.

(3) FREMLIN, J. H., GENT, A. W., PETRIE, D. P. R., WALLIS, P. J., and TOMLIN, S. G.: "Principles of Velocity Modulation," *ibid.*, 1946, **93**, Part IIIA, p. 890.

(4) PEARCE, A. F.: "A Velocity-Modulation Reflection Oscillator for Wavelengths of about 3·2 cm," *ibid.*, 1948, **95**, Part III, p. 415.

(5) COLLARD, J.: "The Enthrakometer, an Instrument for the Measurement of Power in Rectangular Wave Guides," *ibid.*, 1946, **93**, Part IIIA, No. 9, p. 1399.

(6) COLLARD, J., NICOLL, G. R., and LINES, A. W.: "Discrepancies in the Measurement of Microwave Power at Wavelengths below 3 cm," *Proceedings of the Physical Society*, B, 1950, **63**, p. 215.

(7) HAMILTON, D. R., KNIPP, J. K., and KUPER, J. B. H.: "Klystrons and Microwave Triodes" (McGraw-Hill Book Co., 1948).

(8) PEARCE, A. F., and MAYO, B. J.: "The Design of a Reflex-Klystron Oscillator for Frequency Modulation at Centimetre Wavelengths," *Proceedings I.E.E.*, Paper No. 1301 R, April, 1952 (**99**, Part IIIA, p. 445).